Draft Pilot Study Report

May 24, 2007

The TWG members were the primary authors of this document and the TWG did review and comment on this document as it was being developed. However, the TWG did not have time to fully review, discuss, and agree on this document. The TWG agreed to send this document to the FACDQ with the understanding that it is for information and decision-making purposes, but not a fully vetted document.

Pilot Study Report

Table of Contents

Author Initials	Name
BA	Bob Avery
BE	Brian Englert
DR	Dick Reding
JPH	John Phillips
JPL	Jim Pletl
KM	Ken Miller
KO	Ken Osborn
LL	Larry LaFleur
NT	Nan Thomey
RR	Rick Rediske
SB	Steve Bonde
SW	Steve Wendelken
TF	Tim Fitzpatrick
TWG	Technical Work Group
ZE	Zonetta English

I. Pilot Report Executive Summary (DR/JPH/BE)

The FACDQ initially considered over a dozen detection and/or quantitation procedures, and narrowed these down to three pairs that estimate both a detection limit (DL) and a quantitation limit (QL). These three candidate pairs differ from the MDL (a DL) and ML (a QL) approach in several ways. They more fully account for bias, accuracy, temporal variability, and initial and ongoing verification.

Five analytical methods and a total of 55 unique analytes provided a evaluation of chemistries and analytical techniques, each with different precision and accuracy capabilities. The FACDQ decided, for the convenience of the pilot study, to require the same MQO targets for every chemical and analytical method studied, since most of the procedures allowed for some flexibility in the selection of different MQOs. These experiments are performed over several weeks, use blanks and spiked samples which may encompass several different concentrations of the target analyte. The MQOs selected by the FACDQ and TWG for the pilot study were 20% RSD, a 50% to 150% recovery range, and false positive and negative rates of ≤1%. Review of these data made it possible to identify both the strengths and weaknesses of the different DL and QL limit procedures.

In summary, the pilot study was able to successfully evaluate several candidate DL and QL procedures using a substantial number of laboratories and methods. The pilot study findings were very informative and can be used to select and modify one or more DL and QL procedures to meet FACDQ needs. It is important for the FACDQ to take into consideration the application of MQOs and the regulatory use(s) in order to make a final decision regarding the selection and implementation of a procedure.

Pilot Study Observations (KM)

1. The detection and quantitation limit procedures meet most of the MQOs most of the time.

2. The likelihood of a procedure limit meeting the targeted MQOs is heavily affected by what the procedure itself targets (i.e., whether the procedure MQO(s) match the FACDQ MQOs) and how it targets them (i.e., does it estimate the lowest concentration at which the procedure MQO(s) are met, or does it demonstrate that the MQO(s) can be met at a chosen concentration).

3. For some limits (false positive rate for all detection limits, RSD at the IQE20%), it would be expected that the FACDQ MQO would be met half the time, because the procedure targets the lowest concentration to achieve the FACDQ MQO. For others (false negative rate at the IDE), it would be expected that the FACDQ MQO would be met less than half the time because the procedure targets a less stringent MQO than the study. For other limits (mean recovery at the LCMRL), it would be expected that the FACDQ MQO would be met more than half the time because the procedure targets a more stringent MQO than the study, or the procedure does not target the lowest concentration to yield the FACDQ MQO.

4. For quantitation limits, the most difficult MQO to meet is the false negative rate.

   False negative rates were higher when making detection decisions based on detection limit compared to making detection decisions based on instrument signal.

Unlike the other FACDQ MQOs, the false negative rate is based on two limits, the detection and the quantitation limit. As a result, a high false negative rate could be due to a biased-high detection limit or a biased-low quantitation limit.

For ASTM (IDE and IQE) limits (both single lab and interlab), false negative rates were higher when detection based on Lc compared to detection based on Yc. This was especially true for Method 625, for which recoveries were less than 100% throughout the concentration range, and as a result the recovery-corrected Lc was greater than Yc.

5. Mean recovery criterion rarely failed for most analytes/labs, but criterion failures occurred both on the high side (>150%) and the low side (<50%). Low failures generally occurred for a few "problem" analytes (see Method 625, 608 Endosulfans). High failures occurred mostly for Method 300.0.

6. Big differences were observed, both in performance and limit MQO success, between laboratories. Big differences were also observed between analytes. Some laboratory differences were attributable to differences in how method was applied (ex: extraction technique for Method 625).

7. For detection limits especially, variability between lab limits tended to be greater than the variability between the different procedure limits for a single lab, which may affect the assessment of the different procedures. For quantitation limits, the variability between lab limits tended to be greater; however, this was likely due to the different quantitation limits targeting different MQOs.

8. The level of background contamination (or other blank bias) varied widely between laboratories. Blank bias observed in blind samples was not always observed in the existing blanks (or vice versa). This was most frequently the case for Method 300.0.

9. For uncensored methods (chiefly 200.7), the amount of existing blank data varied significantly. Therefore, the precision of the calculated false positive rates differed between laboratories. In addition, the false positive rates of interlaboratory limits were more heavily affected by some labs than others (e.g.: copper).

10. Not all laboratories interpreted the ACIL procedure, and the SOW instructions regarding the ACIL procedure, in the same way. Therefore, some ACIL limits and limit evaluations are more representative of the written ACIL procedure than others.

11. Detection limit procedures based on extrapolation from spiked data are more prone to unexpected false positive rates. Especially for censored methods, this is often due to the recovery vs. concentration relationship not being linear from 0 to quantitation, as assumed by the procedures. In some cases, this was due to choice of spike level; however, the false positive rates and false negative rates were not generally improved by using lower spike levels (see Study Design section - II.d iv.)

12. False positive rate evaluation for uncensored methods is geared toward the ACIL, because data used to calculate limits are also used to determine false positive rates.

13. For Methods 608 and 625 interlaboratory RSDs exceeded 20% throughout the concentration range for several analytes, and therefore the calculated limits would automatically fail this MQO criterion. This was due to low biases of these methods, and differences between laboratories throughout the concentration range.

14. Based on the procedure instructions, software, and pilot study design, some procedures will yield a limit for every analyte/lab, while others will often fail to yield a limit. Therefore, MQO estimates for some procedures summarized over analytes and labs will be more heavily affected by "difficult" analytes (or labs) than MQO estimates for other procedures.

15. Independent of procedure, the order at which the Pilot Study MQOs were met varied by method, analyte and/or lab. For uncensored methods and Method 300.0, the RSD target tended to be met at lower concentrations than the mean recovery rate MQO. However, for Methods 608 and 625, the RSD target tended to be met at higher concentrations. The false negative rate MQO was heavily affected by both the variability and bias observed (and as a result, where the recovery and RSD MQOs are met), but also by the level of blank bias.

Table 1. Detection and Quantitation Limit Procedures to be Evaluated in the FACDQ Study

Name of Procedure	Type of Limit Addressed	Type of Procedure	Description
EPA Office of Ground Water and Drinking Water (OGWDW) Hubaux and Vos Y_c	Detection	Single and Multiple Labs	Additional information about this procedure can be found at http://www.epa.gov/OST/methods/det/faca/techworkgroup/hubaux.pdf
ASTM Interlaboratory Detection Estimate (IDE)	Detection	Interlab	Determines the lowest concentration at which there is 90% confidence that a single measurement from a laboratory selected from the population of qualified laboratories represented in an interlaboratory study will have a true detection probability of at least 95% and a true nondetection probability of at least 99% (when measuring a blank sample.
EPA OGWDW Lowest Concentration-Minimum Reporting Level (LC-MRL)	Quantitation	Single or Multiple Labs	Considers both accuracy and precision in analytical measurement, and is based on linear regression of multiple concentration replicate data and a 99% prediction interval around the regression line. The LC MRL is intended to be used primarily during analytical method development, although laboratories may determine LC MRLs as an aid in determining their single-laboratory minimum quantitation level. A much simpler procedure also was developed to allow laboratories to determine if the minimum reporting level (MRL) they use (either an MRL required by regulation or one set by the laboratory or their client) meets a set of established data quality objectives. Additional information about the LC MRL, including a computer application to calculate LCMRLs can be found at: http://www.epa.gov/OGWDW/methods/sourcalt.html#Mlcmrl.
Interlaboratory Quantitation Estimate (IQE)	Quantitation	Interlab	The IQE is computed to be the lowest concentration for which a single measurement from a laboratory selected from the population of qualified laboratories represented in an interlaboratory study will have an estimated relative standard deviation of (typically) 10, 20, or 30%.
American Council of Independent Laboratories (ACIL) Proposed Procedures for Determining the Method Detection Limit	Detection	Single Lab	Determines method detection limit (MDL), the lowest result that can be reliably distinguished from a blank. Separate sets of steps are provided for two types of methods: those for which a blank consistently generates results and methods for which a blank does not.
American Council of Independent Laboratories (ACIL) Proposed Procedures for Determining the Minimum Level	Quantitation	Single Lab	Determines the minimum level (ML), the lowest level that meets five conditions: • Results from spikes at the ML must be above the MDL. • The ML must be at or above the lowest calibration level (or calibration verification standard for tests with a single-point calibration). • The ML must be at least two times the MDL • The relative standard deviation of results from spikes at the ML must be less than 20%. • The average recovery of spikes at the ML must be within 50-150%. Separate sets of steps are provided for two types of methods: those for which a blank consistently generates results and methods for which a blank does not.

Table 2. Analytical Methods to be Used and Analytes Targeted in the FACDQ Study

Method	Rationale for Selection	Analytes Targeted in Each Method		
EPA Method 200.7, Trace elements via ICP-atomic emission spectroscopy	This is a widely used multi-analyte method using optical techniques to determine metals. Detection limits for this method can be driven by blanks or instrumental sensitivity, and the method is subject to false positives.	Aluminum Antimony Arsenic Barium Beryllium Cadmium Calcium Chromium Cobalt	Copper Iron Lead Magnesium Manganese Molybdenum Nickel Phosphorus Potassium	Selenium Silver Sodium Thallium Vanadium Zinc
EPA Method 300.0, Determination of Inorganic Ions by Ion Chromatography (Method A)	This is a widely used multi-analyte method used to target several stable analytes.	Bromide Chloride Fluoride	Nitrate-N Nitrite-N	Ortho-Phosphate-P Sulfate
EPA Method 335.4, Total Cyanide	This spectrophotometric method is widely used to determine total cyanide.	Total Cyanide		
EPA Method 608, Organochlorine Pesticides and PCBs *(Note: PCB-1016 and 1260 will not be targeted in the Regression Design.)*	This is a widely-used GC/ECD method that targets multi-component analytes and that also can be used to target single component pollutants.	Aldrin Alpha-BHC Beta-BHC Delta-BHC Gamma-BHC Alpha-Chlordane Gamma-Chlordane	4,4'-DDD 4,4'-DDE 4,4'-DDT Dieldrin Endosulfan I Endosulfan II Endosulfan sulfate	Endrin Endrin aldehyde Heptachlor Heptachlor epoxide PCB-1016 PCB-1260
EPA Method 625, Capillary Column Gas Chromatography/Mass Spectrometry ***Note:*** Analytes that are also targeted in EPA Method 608 (e.g., Aroclors, chlordane, toxaphene) are not targeted by Method 625 in the FACDQ study	This is a widely used multi-analyte method using GC/MS techniques to determine semivolatile organic compounds. Detection limits for this method are often driven by qualitative identification criteria; the sample preparation stage of the method can be a source of imprecision.	Acenaphthene Acenaphthylene Anthracene Benzo(a)anthracene Benzo(b)fluoranthene Benzo(k)fluoranthene Benzo(a)pyrene Benzo(ghi)perylene Benzyl butyl phthalate Bis(2-chloroethyl)ether Bis(2-chloroethoxy)methane Bis(2-ethylhexyl)phthalate Bis(2-chloroisopropyl)ether 4-Bromophenyl phenyl ether 2-Chloronaphthalene 4-Chlorophenyl phenyl ether Chrysene Dibenzo(a,h)anthracene	Di-n-butylphthalate Di-n-octylphthalate 1,3-Dichlorobenzene 1,2-Dichlorobenzene 1,4-Dichlorobenzene 3,3'-Dichlorobenzidine Diethyl phthalate Dimethyl phthalate 2,4-Dinitrotoluene 2,6-Dinitrotoluene Fluoranthene Fluorene Hexachlorobenzene Hexachlorobutadiene Hexachloroethane Indeno(1,2,3-cd)pyrene Isophorone Naphthalene	Nitrobenzene N-Nitrosodi-n-propylamine Phenanthrene Pyrene 1,2,4-Trichlorobenzene 4-Chloro-3-methylphenol 2-Chlorophenol 2,4-Dichlorophenol 2,4-Dimethylphenol 2,4-Dinitrophenol 2-Methyl-4,6-dinitrophenol 2-Nitrophenol 4-Nitrophenol Pentachlorophenol Phenol 2,4,6-Trichlorophenol

Fall '07 Test of 3 Single-lab DL/QL Procedures in Several Labs with 5 Analytical Methods

Method	Class Analyte	Number of Labs*	Number of Analytes Analyzed	ACIL DL/QL Limits Calculated **	ASTM Limits Calculated				OGWDW Limits Calculated	
					Yc/Lc	IDE	IQE20	IQE30	HV Yc	LCMRL
EPA Method 200.7	Trace elements via ICP-atomic emission spectroscopy	8	11	88/88	88/88	88/88	88/88	88/88	88/88	85/85
EPA Method 300.0	Determination of Inorganic Ions by Ion Chromatography (Method A)	7	7	45/45	45/45	45/45	45/45	45/45	45/45	42/42
EPA Method 335.4	Total Cyanide Distillation with Semi-Automated Spectrophotometry	7	1	7/7	7/7	7/7	7/7	7/7	7/7	5/6
EPA Method 608	Organochlorine Pesticides and PCBs by Gas Chromatography/Electron Capture Detector	6	18	108/108	108/ 108	106/ 104	101/ 101	102/ 102	108/ 108	60/71
EPA Method 625	Extractable Semivolatiles Capillary Column Gas Chromatography/Mass Spectrometry	7	18	126/126	126/ 126	126/ 126	120/ 122	125/ 126	126/ 126	64/62

*Number of labs which submitted complete data sets. Procedures columns indicate number of labs for which were able to calc DL and QLs using each of the three Procedures.

** Number of limits determined is number calculated without outlier removal/number calculated with outlier removal

Fall '07 Test of 2 Interlab DL/QL Procedures in Several Labs with 5 Analytical Methods

Method	Class Analyte	Number of Labs*	Number of Analytes Analyzed	ASTM Limits Calculated				OGWDW Limits Calculated	
				Yc/Lc	IDE	IQE20	IQE30	HV Yc	LCMRL
EPA Method 200.7	Trace elements via ICP-atomic emission spectroscopy	8	11	11/11	11/11	11/11	11/11	11/11	11/11
EPA Method 300.0	Determination of Inorganic Ions by Ion Chromatography (Method A)	7	7	7/7	7/7	7/7	7/7	7/7	6/6
EPA Method 335.4	Total Cyanide Distillation with Semi-Automated Spectrophotometry	7	1	1/1	1/1	1/1	1/1	1/1	0/1
EPA Method 608	Organochlorine Pesticides and PCBs by Gas Chromatography/Electron Capture Detector	6	18	18/18	17/17	10/12	17/17	18	3/3
EPA Method 625	Extractable Semivolatiles Capillary Column Gas Chromatography/Mass Spectrometry	7	18	18/18	16/16	15/15	16/16	18/18	1/2

*Number of labs which submitted complete data sets. Procedures columns indicate number of labs for which were able to calc DL and QLs using each of the three Procedures.

** Number of limits determined is number calculated without outlier removal/number calculated with outlier removal

II. Technical Details and Full Narrative of Above

a. Review of Goals, Objective, and Limitations (BE/RR)

i. Goals

The goal of the pilot study conducted in late 2006 was to evaluate several detection limit and quantitation limit procedures using five analytical methods for chemicals in at least eight laboratories. Detection and quantitation limit estimates were calculated from these data using three pairs of candidate det-quant procedures. The analytical methods were selected to represent a range of measurement technologies and sample preparation techniques, as well as a mixture of both well behaved and less tractable analytes, to provide a typical yet challenging test for the candidate procedures. Several spike concentrations were selected – some were prepared by each laboratory, others were prepared by a vendor and used by all labs. Measurement quality objectives (MQOs) were set for four parameters: false positive and negative rates (1% each), precision (±20%), and a recovery range of 50% - 150%. This study is described in more detail in the Pilot Study Design document approved by the FACDQ at their July 2006 meeting.

ii. Objective

There are several characteristics the FACDQ would like in a det-quant procedure. An objective of the pilot study was to assess the extent to which these characteristics were demonstrated by each candidate procedure. We assessed this using a set of questions. Some questions were answered during the pilot, some afterwards, some by the participating laboratories, and others by the technical workgroup.

A key question is: "Was the procedure being performed correctly by contract laboratories during the pilot study?" A great variation in the laboratory's performance may be due to a misinterpretation of the written procedure. This might indicate that the procedure is poorly written rather than poorly constructed.

Another question is whether the procedure achieved its intended purpose, which is measured by determining how well it met its objectives. For example:

- Did the IQE20 achieve 20% RSD at the IQE, when it was achievable?
- Did the IDE achieve a 5% false negative error rate at the L_C or Y_C for measurements at the IDE?
- Was the false positive error rate at the ASTM L_C one percent?
- Were the individual laboratory recoveries within the range of 50-150% at the determined LCMRL, when it was achievable?

Yet, another question is whether the procedure met the fixed MQOs established for the pilot study. For example:

- Did the procedure meet the recovery MQO of 50% to 150% at LQ, when it was achievable?
- Did the procedure meet the ± 20% precision MQO at LQ, when it was achievable?
- Did the procedure meet the 1% or lower false positive rate MQO for LC?
- Did the procedure meet the 1% or lower false negative rate MQO at LC for the true value at LD or LQ?

The following are the seven study questions:

1. Is the procedure clearly written?
2. Can the data be easily processed in the laboratory?
3. Was the procedure performed correctly?
4. Does the experimental design unduly influence the outcome of the study? Additional clarifying questions from the Multi-lab Subgroup include:
    a. Type of method (censored, uncensored, etc.)
    b. Works equally well if analyte recoveries are uniformly low, uniformly high, or highly variable
    c. Choice of outlier test (not mandated by procedure?)
    d. Number of different concentrations tested
    e. Number of replicates per concentration tested
    f. Magnitude of concentrations tested
    g. Relative relationship between spikes (0.25x, .5x, x, 2x, 4x, etc.)
    h. Number of laboratories
    i. Number of analysts per study or per laboratory
    a. Number of instruments per study or per laboratory
    b. Sample preparation
    c. Number of different days for which analyses are conducted per laboratory
    d. Time span over which analyses are conducted per laboratory (week, month, quarter, year)
    e. Number of data points per detection or quantitation limit calculation
5. Does the procedure achieve its intended purpose?
6. Does the procedure work for all different types of analytical methods?
7. Does the procedure work if applied to real world sample matrices? (This may also include a broader question evaluating how the procedure applies to real world matrices.)

iii. Limitations

The Pilot Study Design Team and the Technical Work Group agreed early that while there is substantial funding available for this effort, this funding is limited. They understood that there is a limited time period for conducting the pilot and analyzing the data. The pilot study was designed with these conditions in mind. Tradeoffs were considered and evaluated on what could realistically be accomplished and still provide value to the committee.

Some of the tradeoffs that the Design Team and Technical Work Group understood with the present pilot study design include:

- The timeframe for collecting laboratory data is a maximum of 45 calendar days, however labs are required to test over a 15 working day period to collect data, which meant that the pilot study will not provide ongoing, quarterly, or annual verification of results as specified in some of the procedures
- The pilot study will not confirm results by subsequent spiking at the calculated limits, though this could be done in a post-FACDQ pilot
- The pilot study will only use reagent water matrices, though real world matrices could be tested in a post-FACDQ pilot
- Although the pilot study will evaluate five analytical methodologies, there are others that may be desirable for testing in a post-FACDQ pilot

- The pilot study will not capture some sources of intra-lab variability such as variability between analysts and between instruments, though a post-FACDQ pilot could do so.

The technical workgroup in "Features of the Pilot Study and Desired Features of a Post-FACDQ Pilot Study" (June 22, 2006), contrasted what the pilot study was expected to do with what the workgroup recommended future testing do to mitigate the limitations of the 2006 pilot study. Desired features of future procedure testing included testing lab performance over a period of six to twelve months, testing more samples to obtain better estimates of false positive and false negative rates.

b. Review of Pilot Study Design + MMA Design (BE/RR)

i. Pilot Study Design (BE)

The following provides a summary of the basic elements of the pilot study. The pilot study incorporated three separate sets of lab analyses:

- Single-Lab Study Analyses
- Regression Study Analyses
- Aroclor Confirmation Analyses

The purpose of the single-lab study is to determine the MDL and ML following the written ACIL procedure. The purposes of the blind regression study are to calculate single-laboratory and interlaboratory variants of the limits described in the OGWDW and ASTM procedures, and to evaluate whether all limits met the MQOs specified by the FACDQ and in the procedures themselves. Specific differences between single-lab study and regression study are: (1) labs will prepare their spikes and calculate their detection and quantitation limits under the single lab design; (2) single-blind spikes will be sent to labs under the regression design; (3)the Team will calculate a detection and quantitation limit from the results of the analyses of the regression spikes using the ASTM IDE and IQE procedures; and (4) labs that choose to only bid on the two target Aroclors (1016 and 1260) in method 608 will do so only under the single lab design. This exception for Aroclors is made to conserve resources and take advantage of existing Michigan Manufacturers Association (MMA) Aroclor data. These laboratories would also perform the third set of lab analyses, the Aroclor confirmation analyses.

General Pilot Study Design

- Minimum of 8 labs
  - Labs will be solicited for interest and must pre-qualify in order to bid.
  - Pre-qualified labs may bid on one or more methods.
  - The 8 qualified bidders that give the best value to EPA will be selected for each method.
- Five analytical methods
  - EPA Methods 200.7 (metals), 300.0 (nitrate, ions), 335.4 (cyanide),
  - 608 (Pesticides) and 625 (organics).
  - Analytes listed in both 608 and 625 will be analyzed by 608 only.
- Historical blank data collected from labs
  - Analyte data generated during last 30 analytical batches or last 6 months, whichever yields the greater number of results from the instrument(s) used in the study.
  - Data generated on the same instrument will be used in the study.
  - Report blank data without any reporting limit censoring; may require labs to review/revise their historical data.
  - Blanks used in calculation and evaluation of detection limits

Regression Study Design

- A range of concentrations will be analyzed for each method
  - 12 concentrations, including a blank sample.
  - Exact spike levels determined by Team based on lab proposals during the pre-qualification stage (each lab reviewed the LC-MRL procedure and stated which spike levels they would use to perform the procedure; the Team chose spike levels to reflect lab responses as much as possible).
  - Concentrations approximated those needed to determine limits using each procedure.

- Ten replicates at each concentration by each lab for each method
  - Concentrations will be blind to labs.
  - A spiking lab prepared and labeled each sample.
  - Samples based on the study spiking scheme approved by the Team.
- PCB Aroclors were not evaluated for regression-based procedures using new data
  - Existing data with appropriate design is available from MMA PCB dataset.
  - Limits were calculated and confirmed using the same approach that was used to evaluate these limits with pilot study data (to the extent based on the MMA design).
  - While Aroclors were not included in the regression design, additional analyses of two Aroclors at three concentrations will be analyzed for use in confirmation of single-laboratory limits (see below).

Single Lab Design

- One single-laboratory procedure (Modified ACIL - Revision 6.0) was evaluated by each laboratory independently
  - Each laboratory chose initial spike level, prepared samples, analyzed samples and determined the limits.
  - Both start-up (seven replicates) and ongoing limits (based on twenty replicates) were calculated.
- Two Aroclors (1016, 1260) also were analyzed using Method 608 (confirmation based on additional laboratory analyses at multiple concentrations by each laboratory).
- MMA PCB data inappropriate to apply the single-laboratory procedure.
  - Table ?? (Target Analytes for Pilot Study Design) lists the analytes included in the study.

Elements of the Pilot Design

Procedures

As part of their assessment process, the FACDQ identified several candidate procedures for determining detection and quantitation limits and determined that a controlled study of these procedures is necessary to effectively evaluate their merits and limitations. Due to cost considerations, FACDQ members agreed to test only three detection limit procedures and three quantitation limit procedures. These procedures were selected by the FACDQ after prioritization and consideration of several characteristics that were determined to be essential for a successful procedure. In selecting the final set of procedures for testing, the FACDQ chose a pair of detection and quantitation limit procedures that require use of an interlaboratory study to determine the limits (interlaboratory procedures). The FACDQ also chose a pair of procedures that can each be determined through a study of multiple laboratories or in a single laboratory. The final pair of procedures is designed to be determined in a single laboratory only. Table ?? summarizes the procedures selected by the workgroup for evaluation in this study.

Additional procedures, including the Consensus Group Proposed Procedures for Estimating the Critical Level and Quantitation Limit, and the East Bay MUD Procedure for Determining a Detection Limit using Lab QC, were not included in the study. However, some of the Pilot Study data were used to evaluate these procedures, as discussed in Section II c. iii.

Note (1): Although the Hubaux-Vos and LCMRL are single-lab procedures, they will be tested using the spikes developed for the IDE and IQE under the regression design.
Note (2): Detection or quantitation procedures that provide interlaboratory

estimates for detection and quantitation differ from the other procedures in the table that provide intralaboratory estimates.

"Interlab" refers to calculating one limit from the results of experiments in several labs, with that limit calculated based on interlaboratory variability. An example is the IDE/IQE that will be tested in the regression design of the committee pilot study.
"Intralab" or single-laboratory refers to calculating a limit from the results of experiments conducted in one lab. An example of this is the ACIL procedure that will be tested in the single laboratory design of the pilot. Any set of single laboratory limits may be pooled to calculate a multi-lab limit, based on the mean and variability of the different single-lab limits.

Measurement Quality Objectives

The committee set measurement quality objectives (MQOs) for the pilot test at its March 2006 meeting. Each procedure limit would be assessed to determine whether the target MQO characteristic was met. The established MQOs are listed below:

- Mean Recovery within 50-150% (quantitation limits)
- RSD at most 20% (quantitation limits)
- False Negative Rate 1% (quantitation limits)
- False Positive Rate 1% or less detection limits)

Laboratory Analysis Requirements

In order to provide the most neutral of conditions for the pilot, participating laboratories were required to meet the listed conditions:

- All laboratory analyses must be performed on a calibrated instrument.
- Labs will report if they recalibrate during study.
- Labs will follow the calibration requirements in the method.
- All method-specified lab blanks must be analyzed before each batch.
- Reflecting routine analysis, blanks should be as free from contamination as
- possible.
- Labs will follow only relevant method-specific lab quality control
- requirements.
- Samples must be carried through all preparation and laboratory analysis steps
- as are typically used for wastewater samples, such as:
  - Digestion, extractions, and cleanups;
  - Instrument parameter set ups; and
  - Laboratory staff that conduct the study laboratory analyses be the same staff that routinely conduct laboratory analyses by that method.

Data Reporting

In order to evaluate the data effectively the Team identified the following data reporting requirements:

- Labs should not censor any results for which the instrument yields a numeric result. This means the laboratory should report even negative values or values less than the laboratory's current reporting limit.
- Labs should identify any qualitative identification criteria they use that differ from the criteria specified in the EPA method, e.g. criteria used to identify and quantify analytes using GC/MS.

- Labs will report run logs weekly to the prime contractor ("the contractor") to allow monitoring of study status, holding times, and laboratory analysis sequences.
- Labs will report summary level electronic data by week to the contractor, beginning 14 days after completing first week of laboratory analysis, and concurrently provide all supporting raw data. Raw data includes peak areas for calibration data and analytical runs.
- A standardized electronic format will be developed & provided to labs.
- Labs must use this standardized format to expedite data review, data distribution, and data analysis.
- Labs should retain raw data for a period of 5 years and provide it on request (and at additional cost negotiated as necessary).

Michigan Manufacturers Association PCB Study Summary (RRR)

In 2000, the Michigan Manufactures Association (MMA) conducted an interlaboratory study to determine the detection and quantitation limits for PCBs by method 608. The study included the following components:

- Eleven laboratories participated in the study. Laboratories were pre qualified by having least one valid State or Federal certification for the analysis of PCBs by method 608, successfully completed at least two PE samples sets for PCBs by method 608 over the past two years, and having at least 75% of all PE sample results over the past two years within the specified warning limits.
- Study duration of 15 weeks
- DI water matrix with 3 clean-ups required
- 9 samples (1 blank and 8 spiked concentrations) submitted every 1-2 weeks
- Spiked concentrations were submitted in random order as Youden Pairs on alternate weeks. Laboratories were not aware of sample concentration or the identity of the blank.
- Report all positive data using a minimum of 3 aroclor peaks for identification and quantitation.
- The final data set included 5 replicates of each Youden Pair. Data set contained 10 replicates of average Youden Pair result.

The resulting data set meets the minimum requirements to calculate the ASTM IDE/IQE (at least 6 participating laboratories, 5 concentrations, and 5 replicates). The presence of additional laboratories, concentrations, and replicates will allow the data set to be analyzed with respect to the influence of these variables on the IDE/IQE calculations. In addition, the data set is sufficient for the calculation of the LCMRL (minimum of 5 replicates at 4 concentrations) and Hubaux & Vos. If the Youden pairs are averaged, the data set is sufficient to explore how of the number of replicates and concentrations influence the LCMRL/ Hubaux & Vos calculations.

The study also attempted to deal with the problems associated with detection and quantitation limit studies associated with censored methods and positive blank data. The method 608 protocol requires a minimum of 5 peaks to be present for aroclor identification. The MMA study ask laboratories to report all positive data using a minimum of 3 peaks and record the number of peaks used for identification. Positive data from laboratory and study blanks were included in the data set. Although these data were not used to produce blank corrected results, the information is available to further explore the influence of this variable. From these data, the influence of the number of aroclor peaks reported and positive blank data can be evaluated in relationship to the selected detection/quantitation limit estimation methods.

The MMA PCB interlaboratory study provides an approach to deal with the problems associated with the estimation of detection and quantitation limits in censored methods. By exploring the relationship of the number of participating laboratories, concentration levels, and replicates, useful information can be obtained to design future interlaboratory pilot studies. In addition, the study results show the sensitivity of the detection limit estimation procedures to outlier removal and instrument calibration.

c. Limit Calculations & Analysis (KM)

i. – iii. Assumptions and Data Handling Practices

Data Removal

Statistical analyses were performed to assess the effect of data usability decisions on the calculation and confirmation of detection and quantitation limits. Data usability decisions were based on removal of statistical outliers and mis-identified compounds. Outlying laboratories were identified using the Youden outlying laboratory test, and outlying results were identified using Grubbs test. Additional data removal was done for the MMA study based on Aroclor confirmation. One laboratory incorrectly identified Aroclor 1016 as Aroclor 1242 for all analyses, and two laboratories each misidentified one Aroclor 1260 result. (Incorrect identification of target compounds was not observed in the FACDQ Pilot Study.) Statistical analysis were performed in multiple ways as follows, and summarized in Table 1.

MMA data: Statistical analyses were performed with and without removal of outlying laboratories, with and without removal of outlying results, and with and without removal of mis-identified compounds. Grubbs test was run with data combined over all laboratories for all limits.
Pilot Study data: Statistical analyses were performed with and without removal of outlying results. (The outlying laboratory test was not performed due to the smaller number of laboratories performing each method.) Grubbs test was run with data combined over all laboratories for the assessment of interlaboratory limits, but for assessment of single-laboratory limits, Grubbs test was run separately for each laboratory.

See PSR II c i-iii Table 1 Appendix.

Limit Calculation – MMA Study

Single- and interlaboratory detection and quantitation limits were calculated using data from the MMA study. Single-laboratory detection limits included the Hubaux-Vos Yc and single-laboratory variants of the ASTM Yc and Lc (referred to as the SL-Yc and SL-Lc in this document). Single-laboratory quantitation limits included the OGWDW LCMRL and single-laboratory variants of the ASTM IDE and IQE at 20% and 30% RSD (referred to as the SL-IDE, SL-IQE20 and SL-IQE30 in this document). Interlaboratory detection limits included the ASTM Yc and Lc and interlaboratory variants of the HV Yc (referred to as the IL-HV Yc in this document). Interlaboratory quantitation limits included the ASTM IDE, ASTM IQE at 20% and 30% RSD, and interlaboratory variants of the LCMRL (referred to as the IL-LCMRL in this document).

Each detection and quantitation limit procedure evaluated requires use of data from a smaller set of data than were generated in the study. Therefore subsets of the MMA data were generated for limit calculation; these subsets were chosen to best reflect how the written procedures being evaluated would be applied in practice. Charts depicting the subsetting of data for limit calculation using the MMA data are presented in Figures IIc.i - IIc.iv in Pilot Study Appendix.

Interlaboratory limits: Only five concentrations were needed to calculate these limits. These five concentrations were chosen by randomly selecting one of the two concentrations within five of the Youden pairs analyzed in the study. For quantitation limit calculation, the highest and two lowest Youden pairs were excluded from this selection process, and for detection limit calculation, the lowest and two highest Youden pairs were excluded from this selection process.

As a result, the quantitation limits were determined using results of samples spiked at higher spike levels than the results of samples used to calculate detection limits. At each of the randomly selected concentration levels, a single result per laboratory was then selected to create the MMA subsets needed for these calculations.

Single-laboratory limits: All results from each laboratory for the same five randomly selected concentrations that were used to determine the interlaboratory limits were used to determine the single-laboratory limits.

All limits were calculated by CSC (the study coordination contractor) using SAS software. For the ASTM limits, the appropriate standard deviation model was chosen using diagnostic plots and hypothesis tests, as outlined in the IDE and IQE procedures. For the OGWDW limits, the choice of unweighted vs. weighted linear regression was based on the Cook-Weisberg test for constant variance, as outlined in the written LCMRL procedure.

A more detailed discussion in how limits were calculated using the MMA data is presented in the Pilot Study Report Appendix.

Limit Calculation – Pilot Study

Unlike the MMA study, some of the single-laboratory limits were calculated by the participant laboratories in the FACDQ Pilot Study. This was due to a number of reasons. One of the goals of the Pilot Study was to yield information on the ease and understandability of the different procedures in practice; therefore it was beneficial to have the laboratories proceed through the entire procedure as much as possible, including the calculations. Also, some procedures require that some of the calculations be completed in order to choose an appropriate spike level. Therefore, the laboratories were instructed to perform the calculations necessary to determine the ACIL MDL, ACIL ML, and LCMRL, and submit their calculations and results to the study coordination contractor.

For practical reasons, the study coordination contractor (CSC) performed calculation of other limits. For example, interlaboratory limits could not be calculated by individual laboratories because they did not have access to data generated by other laboratories, so CSC performed these calculations. CSC also calculated the single-laboratory variants of the ASTM limits to streamline the study and avoid misunderstandings concerning implementation of a single-laboratory variant of the procedure. The single-laboratory Hubaux-Vos Yc limits were originally intended to be calculated by the laboratories, but due to time and software limitations, these also were ultimately calculated by CSC.

A general discussion of the limit calculations for the different procedures is presented below. More specific information, including charts IIc v. – IIc vii. which depict the subsetting of data for limit calculation, can also be found in the Pilot Study Report Appendix.

ACIL MDL and ML

A copy of the ACIL procedure was provided to each laboratory during the laboratory solicitation process. Once the study began, the laboratories were instructed to determine an ACIL MDL and ML following the written procedure. This included determination of both the initial and ongoing estimates of the MDL and ML. However, some modifications were made to the ACIL procedures in order to accommodate the tighter timeframe involved in the Pilot Study, such that the laboratories were instructed to base their ongoing limits on twenty replicates analyzed over approximately three weeks, rather than over a year, as instructed in the written ACIL procedure.

Once all analyses and calculations were complete, the laboratories provided the calculated ACIL limits and accompanying calculations to the study coordination contractor for review and verification. Any errors made in the calculation process were noted, documented, and corrected where possible.

LCMRL

During the laboratory solicitation process, each laboratory was asked to read and review the LCMRL procedure, and identify the spiking levels (i.e., spiked sample concentrations) they would use to determine an LCMRL for each analyte for each analytical method they would perform in the study. As described in Section II.b of this report, these laboratory-recommended spike levels were compared across laboratories to identify twelve concentrations that would be prepared and used in the study. A spiking vendor was tasked with preparing 10 replicates at each of these 12 concentrations and labeling each sample in such a way that the concentration was blind to the participant laboratories that received them. Once this was completed, each laboratory's five originally selected concentrations were compared to the final study concentrations and matched as closely as possible. The laboratories' originally chosen spiking levels and five study concentrations that were matched to these are presented in the Pilot Study Report Appendix. At each of the five chosen spike levels chosen for a given laboratory, five of the ten replicates were randomly chosen for each analyte.

After laboratories completed analysis of the blind samples, the study coordination contractor notified each laboratory of the sample numbers and corresponding spike levels that should be used to calculate the LCMRL. (These reflected the study concentrations that best matched the laboratory's originally-recommended spike levels.) The laboratories used their own data from the selected spike levels to calculate the LCMRL for each analyte. Each laboratory reported their results (LCMRLs and accompanying calculations) to the study coordination contractor for review and verification. Any errors made in the calculation process were noted, documented, and corrected.

Study Coordinator-Calculated Limits

Single-laboratory detection limits calculated by the study coordination contractor included the Hubaux-Vos Yc, SL-Yc and SL-Lc. Single-laboratory quantitation limits calculated by the study coordination contractor included the SL-IDE, SL-IQE20 and SL-IQE30. Each of these limits was determined following the written procedures, and following the same general framework as used for the MMA data.

All interlaboratory detection and quantitation limits, including the ASTM limits and interlaboratory variants of the Hubaux-Vos and LCMRL, were determined by the study coordination contractor. Limit calculations were performed using one replicate per laboratory and spike level, with spike levels chosen based on the different LCMRL spike level choices submitted by the laboratories for the analyte. Spike levels used to calculate the interlaboratory detection limits were lower than those used to calculate the interlaboratory quantitation limits. The limit calculations followed the same general framework as used to determine the single-laboratory versions of the given procedure and limit.

Limit Confirmation using The MMA and Pilot Study Datasets

Limit confirmation was performed by assessing whether the pre-established measurement quality objectives (MQOs) were met for each of the determined detection and quantitation limits. These MQOs included:

- False positive rates of 1% or below at the detection limits,
- False negative rates of 1% or below at the quantitation limits,
- Mean recoveries between 50% and 150% at the quantitation limits,
- RSD at 20% or below at the quantitation limits, and
- Standard deviation of recoveries at 20% or below at the quantitation limits.

Limit confirmation was performed for each detection and quantitation procedure on both the MMA and Pilot Study data. For Methods 625 and 200.7, limit confirmation analyses were performed on a subset of the laboratory-analyzed analytes only, which are listed in the Pilot Study Report Appendix. For Method 625, 18 of the 52 analytes were included in limit confirmation analyses, and for Method 200.7, 11 of the 24 analytes were included in limit confirmation analyses. This subset of analytes was chosen to accurately reflect the range of method performance among all analytes included in the methods (i.e., including some analytes prone to contamination, some prone to interference, some that are more straightforward to analyze, etc.). For Methods 608 and 300.0, all analytes were included in limit confirmation analyses. However, the procedures used to confirm the limits for Aroclors by Method 608 differed from those used to confirm limits for all other analyses. This was necessary because of differences in the Pilot study design that were intended to allow use of the MMA data set for Aroclors.

With the exception of the Pilot Study Aroclors, which are described at the end of this section, the same general framework was used to compare the calculated limits to the MQOs for both datasets, and is described below. More specific details on these analyses can be found in the Pilot Study Report Appendix.

Confirmation of Detection Limit MQO for All Analytes

All detection limits were evaluated against the MQO that detection limits should yield false positive rates of 1% or less. A false positive was defined as a determination that the analyte of interest was present in a sample, when in fact the analyte was not present. The false positive rates for the different calculated limits were determined using blank sample results.

The amount of blank data available to assess the false positive rate differed between studies. In the Pilot Study, the laboratories were required to submit existing blank data from approximately the last six months for each analyte. In addition, ten blind unspiked samples were included in the study for each analyte, and each laboratory analyzed QC blanks (including calibration and preparation blanks) as required by the different methods included in the study. These additional blanks were included in the false positive rate assessments. In the MMA Study, there were ten blind unspiked samples included in the study, but no other blank data were available.

Each blank result was compared to each of the individual single-laboratory detection limits, and was categorized as a false positive if the result exceeded the given limit. The false positive rate was then calculated as the percent of results exceeding that limit for each analyte, laboratory, and detection limit. False positive rates were determined for the interlaboratory detection limits following the same process, but with blank results from all laboratories used in the assessment.

Confirmation of Quantitation Limit MQOs for All Analytes, Except Pilot Study Aroclors

Unlike detection limits, for which results from one spike level would be used to assess the MQO criteria for all limits, assessing the MQO criteria applicable to quantitation limits requires examining that MQO characteristic at multiple concentrations. This was performed using two different approaches: linear interpolation and nonlinear modeling. Linear interpolation was

performed by assessing the calculated MQO characteristic at the two spike levels most closely surrounding a determined limit, and modeling was performed using the calculated MQO characteristic at all non-zero spike levels in the study. The same general approach was followed for both single- and interlaboratory limits, with only the data from the given laboratory used in the interpolation and modeling of single-laboratory limits, and data from all laboratories used in the interpolation and modeling of interlaboratory limits.

Details on the confirmation of the specific quantitation limit MQO criteria are presented below.

False Negative Rate

A false negative was defined as a determination that the analyte of interest was not present in a sample, when in fact the analyte was present. The false negative rate depends on how much of that analyte actually is present in the sample and, therefore, can be calculated at each of the spike levels included in the study.

When assessing the false negative rate for a given quantitation limit, the method of making the detection decision must be specified. This was done by linking each quantitation limit to an associated detection limit (i.e., the ACIL ML was linked to the ACIL MDL, the LCMRL was linked to the Hubaux-Vos Yc, and the ASTM SL-IQE20, SL-IQE30, and SL-IDE were each linked to both the ASTM SL-Yc and ASTM SL-Lc). A similar method was used to link the interlaboratory quantitation and detection limits. For the three censored methods included in the study (Methods 608, 625 and 300.0), the detection decision was also made based on whether an instrument signal was attained, and this detection approach was applied to all of the single-laboratory and interlaboratory quantitation limits. Each spiked-sample result was categorized as a detect or nondetect based on each of the different detection limits or instrument threshold. The false negative rate at each spike level was then determined as the proportion of samples not detected, based on the specified detection decision approach. These false negative rates were then modeled and linearly interpolated, and the false negative rate at each quantitation limit was determined using the model and interpolation for the corresponding detection approach(es). Because an individual result could be categorized as a detect based on one detection limit and as a non-detect based on another detection limit, a different model was fit for each of the different detection limits.

Mean Recovery and RSD

The mean recovery and RSD were calculated at each of the non-zero spike levels included in the Pilot or MMA studies. The mean recoveries and RSDs were then modeled and interpolated for each of the quantitation limits. Multiple nonlinear models were fit for the mean recovery and RSD for each analyte and laboratory, and the most appropriate model was chosen based on various factors. Unlike the false negative rate determinations, the same model could be used to estimate the mean recovery for each of the single-laboratory limits determined for a given analyte and laboratory, and the same model can be used to estimate the mean recovery for each of the interlaboratory limits determined for a given analyte. Similarly, the same model could be used to estimate the RSD for each of the single-laboratory limits determined for a given analyte and laboratory, and the same model can be used to estimate the RSD for each of the interlaboratory limits determined for a given analyte.

Standard Deviation of Recoveries

Similarly to the mean recovery and RSD, the standard deviation of recoveries was calculated at each of the non-zero spike levels, and was linear interpolated linearly for each quantitation limit.

However, because the standard deviation of recoveries is a function of the mean recovery and RSD, it was not necessary to model the standard deviations separately. The standard deviation of recoveries was instead calculated as the model RSD estimate multiplied by the model mean recovery estimate divided by 100% for each limit.

MQO Limit Summaries

While a comparison of the calculated limits would reflect analyte and laboratory differences rather than the performance of the procedures, a comparison of the MQO characteristics calculated at the individual laboratories can give an indication of how well the procedures generally meet the study MQOs. Therefore, summary statistics of the individual MQO characteristic estimates were calculated to better assess the performance of the various limits for each method. These statistics include the mean and median of all of the individual mean recovery, RSD, standard deviation, and false positive and false negative rate estimates over all laboratories and analytes for each single-laboratory limit, and the median and mean of all of the individual mean recovery, RSD, standard deviation, and false positive and false negative rate estimates over all analytes for each interlaboratory limit.

Limit Confirmation for Method 608 Aroclors

To avoid overlap with the MMA study data, blind Pilot Study samples for Method 608 did not include Aroclors. However, because the ACIL MDL and ML could not be determined using MMA data, laboratories did determine these limits as part of the ACIL sample evaluation in the Pilot Study. Additional samples were necessary to assess whether the resulting ACIL Aroclor limits met the study MQOs; the laboratories were asked to prepare and analyze additional replicate samples for this purpose. For each Aroclor, three sets of five replicates were analyzed, with one set spiked at the laboratory's determined ACIL ML, one set spiked at two times below the laboratory ML, and one set spiked at two times above the laboratory ML.

Because laboratories also submitted existing blank results for Aroclors, the false positive rates at the ACIL MDL were assessed similarly to those for other methods and analytes. However, due to the limited nature of the spiked sample Aroclor results, modeling and interpolation were not performed to assess the quantitation limit MQOs. Instead, descriptive statistics, including the mean recovery, RSD, and minimum concentration, were calculated at each spike level for each laboratory, and compared to the study MQO criteria.

iv. Other calculations or analysis.

1. Lab QC procedure (KO)

Pilot Study Report - Section II C, iii 1 LabQC Procedure 200.7 Metals Data

The Laboratory QC procedure was not included in the pilot study for the calculation of detection and quantitation limits (DLs/QLs). Data have been made available for testing those procedures not originally in the pilot study. Detection limit calculations using the Laboratory QC procedure with the pilot study collected single-laboratory data for metals by EPA 200.7 are discussed in this report.

Background for Laboratory QC Procedure

The Laboratory QC (Lab QC) procedure uses routine QC samples referred to as False Negative Control Samples (FNQS) prepared and analyzed as routine QC samples with each analytical batch. The FNQS is a QC control set at a concentration of two to five times the detection limit

and preferably no more than twice the reporting limit. Using an FNQS approach meets the objectives of 1) providing on-going confirmation of analytical capability in the region of detection, 2) establishing a routine check for false negatives, 3) collecting real-time data representing day-to-day variance for the re-determination of detection limits without the need for a "bench" study, and 4) generating detection limits that are no more than one-half regulatory reporting limits.

Calculating Lab QC RLc Values

RLc is the detection limit expressed as a Critical Value calculated from the variance in FNQS recoveries using the equation:

$$RLc = t*S(Rec)*Median(FNQS)/Median(Rec)$$

Where t = Student's t value (n-1 degrees of freedom and alpha = 0.01), S(Rec) = standard deviation of FNQS recovery, Median(FNQS) = median FNQS value, Median(Rec) = median FNQS recovery.

Recovery is used to compensate for bimodal distributions that can result with small changes in the concentration of replacement standards. The median FNQS is used rather than the average to better represent the center of the distribution and reduce the need to censor for outliers.

Terms and Definitions

Terms used in the pilot study spreadsheets developed by Ken Miller were retained. These terms are reproduced here together with the Lab QC specific terms and definitions in Table 1.

See Table I: Lab QC Calculation Terms and Definitions in Appendices.

California Minimum Levels (CAML) were used as example reporting limits for diagnostic evaluations.

Calculations

Calculations for EPA 200.7 metals with NPDES CAML values are summarized in Table II. RLc values are included for both censored and uncensored data. Values were censored for recoveries outside the limits of 50-150%. The CAML DQO was met for all metals for at least one of the tested concentrations. All concentrations set at more than twice the CAML failed the CAML DQO of achieving an RLc no more than half the CAML.

See Table II: RLc Values Calculated for EPA 200.7 Metals with California Minimum Level Reporting Levels in Appendices.

Diagnostics and Outcomes- Reporting Limit DQO

FNQS diagnostics are used to fine tune an initial set of concentrations. Once established, concentrations need not be revised unless there is a change to the underlying methodology or instrumentation that would change the variance of the method. Diagnostics include the ratio of FNQS concentration to determined RLC, ratio of FNQS concentration to reporting limit (e.g., CAML), and relative standard deviation of FNQS recovery. The outcomes desired include RLc

values that protect against false positives in method blanks, false exceedences of regulatory reporting limits, and false negatives.

Diagnostics as applied to the CAML associated metals are in Table III and Table IV. All instances of RLc values not meeting the CAML DQO were generated from initial spiking concentrations that exceeded twice the CAML values (Table III). All spiking levels less than twice the CAML generated RLc values that met the CAML DQO and some spiking levels exceeding twice the CAML produced acceptable RLc values (Table IV). Neither the RLc to spiking concentration ratio (RLc RATIO) nor the recovery relative standard deviation (RSD FLAG) were predictive of achieving the CAML DQO.

See Table III: RLc Values Not Meeting CAML DQO Requirements and Diagnostics in Appendices.

See Table IV: RLc Values Meeting CAML DQO Requirements and Diagnostics in Appendices.

Diagnostics and Outcomes – False Negatives

The FNQS protects against false negatives if the concentration is set properly. Setting the concentration too high yields detection limits that may exceed the reporting limit DQO; setting the concentration too low can result in false negatives. The RSD flag is set to 25% RSD. The RSD for a normally distributed data set is 40% (for N = 20, t = 2.5, 100/t = 40%). The RSD for a normally distributed data set with an average concentration twice the detection limit will be approximately 25% depending on the relationship between concentration and RSD. 1 Table V summarizes the relationships between RSD Flag and high false negative rates. All RSD Flag values of "OK" or "LOW" were associated with false negative rates of zero and are not included in this table. There was no correlation of false negative rates with either the spike to RLc ratio or the reporting limit DQO.

See Table V: False Negative Rates and RSD Flag in Appendices.

Conclusions

The Lab QC procedure was used to calculate detection limits with the Pilot Study data. The calculated limits were then evaluated against a set of diagnostics for reporting limit DQO, protection against false negatives, and FNQS optimization.
Detection limits (RLc) failing to meet the reporting limit DQO all failed the CAML spiking level test. Spiking levels for FNQS set at concentrations more than twice the reporting limit have a high probability of exceeding one-half the reporting limit.

The ratio of the spiking concentration to the determined RLc is not predictive of passing the reporting limit DQO requirement. High spiking levels exceeding ten times RLc were associated with low RSD (<10%), indicating that the spiking concentration could be decreased if the objective were a lower RLc.

High recovery RSD values were correlated with a high percentage of false negatives.

For methods using the Lab QC approach, a lack of prior experience for setting the appropriate starting concentration would set the upper bound for the FNQS concentration at no more than

1 The RSD at twice detection given the RSD at detection is based on the Rocke-Lorenzato relationship where S(FNQS) ~ SQRT{S(0)^2 + rsd^2(method)}. Calculation specifics available from author.

twice the regulatory reporting limit. The lower bound would be determined after an initial run of three FNQS samples to confirm that the FNQS recovery RSD did not exceed 25%. For methods with pre-existing detection limits, the optimum FNQS concentration would be 2-3 times the detection limit assuming this does not exceed twice the reporting limit. It the RL is exceeded, the method may not be capable of both meeting the reporting limit DQO and protecting against false negatives.

ERRATA: LAB 8, SE RLc(Censored) = 40.3 (not 72.6) 3/8/2007

1. Ruggedness testing (Drinking Water) (TWG)
2. LCMRL, anomalies and calculator (SW)

During data analysis it became evident that an improper LCMRL value was calculated or that an LCMRL value was not calculated at all in many cases. Upon investigation, it was found that specific guidance was missing from the document that was given to users of the LCMRL calculator. This guidance should have addressed two separate issues related to calculator use.

The first case is where the calculator reports a LCMRL value that is equal to the lowest spike level. If this occurs the user should have been instructed to analyze an additional set of replicates at a concentration less than the LCMRL. This process should be continued until the LCMRL value is bracketed by standards. In the pilot study data set there was a total of 87 cases where this occurred. Table 1 gives a breakdown of the number of cases and the remedial action that was necessary to calculate a valid LCMRL.

See Table 1 in Appendices.

When the data set had spike levels less than the LCMRL value available they were used to bracket the LCMRL and generate a valid limit. Of the 87 cases found this procedure corrected 73 of them. In the remaining cases lower spike levels were not present or data quality was poor and valid limits could not be calculated.

In the second case the LCMRL calculator returned an error message stating that an LCMRL could not be calculated. This occurred 119 times when processing the pilot data. In these cases the data quality (bias, precision) at the chosen spike levels (or of the entire data set in some cases) was insufficient to calculate the LCMRL. Of the cases found, adding higher spike levels from the pilot data corrected 42 of them. In the remaining cases, the data quality was so poor that an LCMRL could not be calculated using any combination of spike levels. Table 2 gives a breakdown of the number of cases and the remedial action that was necessary to calculate a valid LCMRL.

See Table 2 in Appendices.

When the LCMRL did not return a value for a data set, it was not because the LCMRL calculator did not work, but rather because it would not report a value when data was not in control and did not meet criteria set forth by the LCMRL. This is essentially a feature that alerts the user to the data quality problem and prevents them from blindly calculating a value that is not reliable.

An LCMRL value was found for all analytes in Methods 300.0. For Method 200.7, the only LCMRL value not found was for Silver, Lab 3, where the highest two levels had a known error, the silver precipitated and recovery levels to fell to 25%. Not surprisingly, this data did not meet

the QC criteria that the LCMRL procedure required. There were 76 data sets that did not meet quality control requirements of LCMRL.

a. One set silver, Lab# 3, had aberrant data caused by the analyte precipitating at the two highest concentration levels, so that recoveries were only 25%. This data was included by the pilot study as data for the LCMRL procedure. The LCMRL did not determine a quantitation level because the data did not meet QC requirements.
b. 53 data sets had LCMRL prediction intervals that fell below the lower QC limit. Fourteen data sets had average recoveries below 50%.
c. 1 data set had LCMRL prediction intervals that were greater than the upper QC limit.
d. 19 data sets had LCMRL prediction intervals that exceeded both QC limits.
e. Two data sets, endosulfan II of Lab# 32 and 3,3'-dichlorobenzidine of Lab# 43, had non-detects at the highest concentration level that was available to use.
f. Four data sets met LCMRL QC requirements once an outlier was removed

Considering that 74 of 76 compounds that failed to meet QC requirements were for organic compounds and that the primary failure was due to recoveries that were biased low, one might conclude that there might be an issue with preservation of shipped samples or extraction procedures.

The use of 25 samples for the LCMRL determination was less than the LCMRL procedure recommends, which is 28. Given limited resources, the need to cut back on collected samples was understandable, but the pilot study actually collected ten samples per concentration level and used only five of these for the LCMRL determination. It is recommended that all samples be used to calculate the LCMRL, and ASTM IDE/IEQ and these estimates compared to pooled multi-lab estimates using all data.

Consensus Group Procedure

In order to evaluate whether how the Consensus Group (GC) procedure might perform, if it had been included in the pilot study, some calculations allowed us to evaluate two of the key differences between the ACIL procedure and the CG procedures. Since the percentage of numeric results for performing the uncensored procedure differs between the two procedures, the percentage of uncensored results in the ACIL single-lab data set was calculated and the uncensored limits calculated following the CG procedure (PSR.II.d.x.Percentage of Numeric Results Difference between ACIL and CG Procedures).

The second key difference between the ACIL and CG procedures was the acceptance requirements for the QL, since the CG procedure had an additional criteria of $s/Lq \times 100 < 20\%$. In order to evaluate this difference s/Lq x100 was calculated for all ACIL single-lab analyte/method combinations, however the spike bias never exceeded 20%.

Intermittent Blank Contamination Guidance

One of the questions posed in the Procedures Report and also commented on several times in this Report is that a procedure to effectively deal with intermittent blank contamination might improve the performance of the procedures. To evaluate this all task one blank data was evaluated for intermittent blank contamination following the FACDQ TWG Draft Intermittent Blank Contamination Guidance, (Draft Intermittent Blank Contamination Guidance). Each MDLs and ML limits generated using this procedure was then compared against the ACIL MDL and ML limits. Percentage false positive and false negative error rate was also calculated using a

direct comparison technique and these results were compared to the ACIL false positive and false negative error rates.

d. Results

i. Is the Procedure Clearly Written? (KM)

Clarity of the Procedures and Suggested Areas of Improvement

All participating labs were asked to submit a written narrative report with each data package. As part of these narratives, labs were asked to comment on the clarity of the procedures and suggested areas of improvement. Most, but not all, labs that participated in the study answered these questions. Their answers are summarized below. *Note:* each laboratory's comments were counted only once, even if that laboratory performed more than one method in the study.

Clarity of the ACIL Procedures

Sixteen labs commented on the clarity of the ACIL procedures.

- Fourteen labs stated that the procedures were written fairly clearly, although some also offered specific areas of improvement and expressed concerns about certain aspects of the procedures. (These concerns and suggested areas of improvement are noted below.)
- One lab commented that the ACIL procedure was not clearly written and that it was difficult to discern which equation should be used for various steps.
- One lab expressed concern about the use of the term "Minimum Level" (ML) instead of "Limit of Quantitation" (LOQ). This lab suggested standardizing on the term LOQ in order to avoid the confusion that results from too many terms representing the same factors.

While the majority of labs stated that the procedure was clearly written, not all laboratories interpreted the procedure in the same way. Multiple laboratories set the ACIL ML equal to two times the ACIL MDL, or the lowest expected result determined in the ACIL procedure, rather than the spike level used to assess the MQOs, as instructed in the procedure. Other laboratories did not choose the spike level at a level appropriate to the ACIL procedure instructions; for example, several labs did not spike at a level at or above the determined ACIL MDL for uncensored methods.

Specific Comments on the ACIL Procedures

Several labs offered specific comments on the procedures and offered suggestions for improvement. Specific comments are listed in The Pilot Study Report Appendix, and are summarized below.

- Most labs felt that the ACIL procedures could be implemented by staff possessing standard lab skills and a basic working knowledge of Excel or other commercial spreadsheet programs. However, some labs also commented that a basic understanding of statistical methods would be necessary to ensure appropriate application of the procedures, and that example calculations or software would be helpful.

- A few laboratories expressed concerns that the ACIL procedure may produce elevated or highly variable limits in some cases. Possible reasons cited by the labs include the use of method blank data to generate an MDL for uncensored methods, requirements to meet the specified precision and accuracy criteria, especially for "poor performers," the lack of a criteria check on setting the ML too high, the ease of meeting the ±50% recovery criterion for the MDL, and requirements to use blanks from the past 20 – 100 analytical runs without censoring 'inappropriate data.'

- The ACIL procedure may be difficult to apply for multi-analyte methods. Many labs expressed concern that a multi-analyte method may require them to fortify the quarterly verification check at several concentrations to address the spike levels for all compounds.

- Concern was expressed about the need to maintain multiple MDLs for each instrument and the need to provide some consistency in reported data for at least a year.

Additional comments on specific sections of the ACIL procedure are presented in Pilot Study Appendix.

Clarity of the LCMRL Procedure

Eleven labs commented on the clarity of the LCMRL procedures. It should be noted that, because the LCMRL procedure was evaluated using blind spiked data rather than lab-spiked data, the laboratories only performed the calculation portion of the procedure. The laboratory comments reflected this limitation of the study design.

- Most reported that they found the LCMRL procedure to be clear and relatively easy to follow and relatively easy to perform using the software provided, but most also offered specific areas for improvement and expressed concern about either the procedure itself of the automated calculator. (These concerns and suggested areas of improvement are presented in The Pilot Study Report Appendix.)
- Labs that had the resources to do so were able to use a computer programmer to format the data in a manner that allowed it to be imported directly into the LCMRL calculator. Labs that did not have such resources reported that the manual data entry required was very time-consuming and error-prone, particularly for the methods with multiple analytes.
- Two labs reported that they found the procedures for determining the LCMRL to be difficult to comprehend. One of these labs added that they found the examples to be helpful in interpreting and evaluating the calculations and graphs, but felt that the procedure lacked direction on how best to produce data for these analyses.
- One lab suggested that if it was important that replicate analyses be analyzed on non-consecutive days, as was required in the study, this requirement should be described in the LCMRL procedure.
- Nearly all of the labs agreed that basic computer and spreadsheet data entry skills were required to implement the procedure. Multiple labs also commented that a chemist and a working knowledge of statistics are needed to calculate and evaluate the resulting limit.

Specific Comments on the LCMRL Procedure

Many labs expressed concerns about the LCMRL procedures and software and offered suggestions for improvement. Specific comments are listed in Pilot Study Appendix, and are summarized below.:

- Labs offered comments on the LCMRL calculator software.
- Labs expressed concerns on how the choice of spike level would affect the resulting limit.
- Several labs expressed concern that the number of replicates at various concentration levels required for procedure could pose a burden for commercial environmental labs

ii. Can the data be easily processed in the laboratory? (BE/KM)

Labs generally stated that the LCMRL calculator was very helpful in performing the required calculations. It was additionally stated that the calculations would be very difficult to complete

without use of the software. Labs which were able to use a computer programmer reported that this assisted in the formatting of data in a manner that allowed it to be imported directly into the LCMRL calculator. Data entry into the LCMRL calculator was other wise time-consuming and prone to error, especially when dealing with multiple analyte methods. Multiple laboratories also stated that the volume of data required in the LCMRL procedure was greater than what is currently required, and would place a greater burden on the laboratories.

The majority of labs stated that only standard lab skills with a basic working knowledge of Excel or other spreadsheet programs were needed to process data using the ACIL procedure. Some labs indicated that a basic understanding of statistical methods aided in the processing of data under the ACIL. Laboratories generally stated that the volume and type of data required in the ACIL procedure was appropriate and manageable.

The amount of data, and the ease at which the laboratory can process the data, would depend on the study design. ASTM IDE and IQE calculations were not intended to be performed by the laboratories themselves.

iii. Was the procedure performed correctly? (JPH/BE/KM)

Not all laboratories involved in the pilot study performed the ACIL procedure in the same way. For uncensored methods, some laboratories did not follow the procedure requirement of choosing an initial spike level at least two times greater than the calculated MDL. As a result, these laboratories were more likely to have high false negative rates at their ML. Additionally, some laboratories did not adjust the rerun startup analyses, despite the initial startup replicate analyses failing one or more of the target MQOs. As a result, the ACIL MLs for these labs would be biased low. Other laboratories did not set the ML to their final spike level, instead setting to two times the MDL or some other calculated value. While these incorrectly calculated ACIL MLs were fixed for the Pilot Study, these errors could also occur when routinely performing the procedure.

While the LCMRL procedure could for the most part be performed correctly, data entry into the LCMRL calculator was prone to error without the aid of a computer programmer for formatting the data. Other laboratories indicated that when using the LCMRL, it was difficult to determine which equations should be used and that replicate analyses need to be performed on non-consecutive days for the procedure to function correctly. This requirement existed in the study but is not specified in the LCMRL procedure.

ASTM IDE and IQE

The ASTM IDE and IQE procedures were performed by Computer Science Corporation (CSC) using a SAS program to carry out the calculations. The IDE and IQE were performed as completely separate procedures using a unique set of data for each procedure. While the ASTM procedures were written as two stand alone procedures, this was primarily a result of the ASTM D19.02 stepwise approach in developing the procedures. When the ASTM DQCALC software was developed it was designed to incorporate D2777 outlier removal options as well as the D6901 (IDE) and D6512 (IQE) standards, so that the IDE and IQE could be determined simultaneously using the same set of data.

To determine if the IDE and IQE were performed correctly using the CSC SAS program, random datasets representing one or more analytical methods were also run using the official ASTM DQCALC software. Results were generated for the LC, IDE, IQE20 and IQE30 and compared to

the SAS results. Results were identical to four significant figures for all datasets evaluated. The YC is not an output of the DQCALC software, however it can be estimated graphically and these results also compared favorably to the SAS results.

A study manager could combine all data when using the ASTM DQCACL software so that more power could be used to generate the IDE and IQE estimates. In general we would expect the more comprehensive dataset to provide better estimates. So to evaluate the robustness of this technique data from both the "idelimitcalconly.xls" and "iqelimitcalconly.xls" files were combined and used to calculate estimates using the DQCALC software. These results for selected analytes are presented in attachment, "PSR.II.d.iii.(Comparison of ASTM Calculations).xls". When the SAS and DQCALC estimates differed slightly the LC and IDE estimates were nearly always lower and the IQE estimates were more often higher using the entire set of data. Overall comparisons calculating the IDE and IQE separately and simultaneously produced remarkably similar estimates for all datasets evaluated.

iv. How did or will the experimental design influence the outcome of the study? (KM)

The Effect of Experimental Design on the Outcome of the Study

As discussed in Section II c., a subset of the Pilot Study data was chosen to best reflect how the particular detection or quantitation limit procedure would be applied in practice. This choice encompassed the number and range of spike levels, and the number of replicates per spike level. However, not all applications of the procedures would follow the same design, and therefore it is of interest to assess how the chosen experimental design affected the outcome of the study, including the ability of limits to achieve the study MQOs, the variability between limits, and the limits themselves. To test this, two alternate limit calculation scenarios were devised. A second experimental design choice, whether or not to remove outliers identified based on a statistical test, was evaluated by performing analyses with and without outlier removal.

Spike Levels for Single-Laboratory Detection Limits

For single-laboratory ASTM and OGWDW procedures, the same data were used to calculate both detection and quantitation limits. While this is consistent with software designed to calculate detection and quantitation limits, it is not necessarily consistent with the written procedures (i.e.,the IDE and IQE procedures do not suggest using the same data, though these are interlaboratory procedures). As the choice of spike levels that were used to calculate single-laboratory limits was based on the laboratories' LCMRL designs and, therefore, was specific to quantitation limits, the detection limits determined from these data may be biased.

To test this assertion, the single-laboratory ASTM and OGWDW detection limits were recalculated using lower spike levels. For each analyte and laboratory, the alternative spike levels were determined by adjusting the original lab-chosen spike levels downward by two levels. For example, if a single-lab limit was originally calculated based on the 4^{th}, 5^{th}, 6^{th}, 7^{th} and 8^{th} lowest spike levels, the limit was recalculated based on the 2^{nd}, 3^{rd}, 4^{th}, 5^{th} and 6^{th} lowest levels. If the limit had originally been calculated with data starting with the 2^{nd} lowest level, the limits were only adjusted downward by one step (i.e., starting with the lowest spike level). The resulting alternative limits were then compared to the original limits by calculating a percent difference (i.e., the adjusted limit minus the original limit, divided by the average of the two limits, expressed as a percent). False positive and false negative rates were also determined using the alternative limits.

In most cases, the alternative limit was lower than the originally calculated limit, with the exception of Method 300.0 (see Table II d. iv. a in Appendix). The effect of the adjustment also tended to be consistent between the different limits. The exception to this observation again occurred in Method 300.0, for which the alternative SL-Y_C and HV-Y_C values tended to be higher than the original limits, while the alternative SL-L_C values tended to be lower than the original limits. This difference was likely due to the high recoveries observed for Method 300.0; the lower spike levels for this method exhibited higher bias and comparable variability compared to the higher spike levels, which resulted in higher limits. The SL-L_c calculated based on the lower spike levels for Method 300.0 was, however, lower than the original limit due to the recovery correction included in that limit calculation.

While the alternative single-lab detection limits tended to be lower than the original limits for most methods, the false positive rates based on these limits did not differ greatly, as shown in Table II d iv b in Appendix. For nearly all methods where the mean or median false positive rate determined using the original limits exceeded 1%, the mean or median false positive rate determined using the alternative limits also exceeded 1%. The largest change in mean false positive rates occurred for Method 625 for the SL-L_C, which was 0.99% for the original limits and 3.5% for the alternative limits. As expected based on the increased limits, the largest drop in mean false positive rates occurred for the SL-Y_C and HV-Y_C for Method 300.0.

Even though the SL-Yc and SL-Lc values tended to decrease when calculated using lower spike levels, the false negative rates based on these limits tended to increase (see Table II d iv c in Appendix). This is quite counterintuitive; because the quantitation limits are not changing, a decrease in the detection limit would be expected to increase the difference between the quantitation and detection limits, and thereby decrease the false negative rates. For the majority of cases in which the detection limit decreased when recalculated, the false negative rate was already 0% (when determined using the original, higher detection limit). However, when the detection limit increased when recalculated, the false negative rate often increased by a large amount. This also occurred for the LCMRL/HV-Yc for total cyanide. Additionally, the difference in false negative rates based on the ASTM SL-Yc and SL-Lc decreased when determined for the alternative limits. For the methods that tended to exhibit high-biased recoveries at low levels (300.0, 335.4, some analytes for 200.7), the false negative rates based on SL-Lc tended to be lower than rates on based on SL-Yc.

Examples of the effect of the lower spike levels on the false positive and negative rates at the calculated single-laboratory detection limits are presented in Figures II d iv a and II d iv b in Appendix.

Number of Spike Levels and Replicates Used in Limit Calculation

The single-laboratory and interlaboratory limits that were originally calculated from the blind sample data utilized only a small subset of the samples analyzed in the Pilot Study. As a comparison, the OGWDW and ASTM single-laboratory limits were calculated using all sample results for the given analyte and lab. Similarly, the OGWDW and ASTM interlaboratory limits were calculated using all sample results for the given analyte. By including all sample results, the alternative limits would differ from the original limits in the following ways:

- The number of spike levels and, therefore, the number of standard deviations and means modeled in the ASTM procedures would be increased.
- The range of spike levels would be broader, making it more likely that both the ranges of constant and increasing variability would be included.

- The larger number of replicates and concentrations would result in lower prediction and tolerance limit multipliers for the detection limits and for the IDE.
- The level of temporal variability would be greater for the interlaboratory limits that were originally determined from only one of the three shipment batches (i.e., one week of analysis).

Similarly to the prior comparison, percent differences were calculated as the alternative limit minus the original limit, divided by the average of the two limits, expressed as a percent. These percent differences are summarized in Tables II d iv d (single-lab limits) and II d iv e (interlab limits) in the Appendix. The alternative limits were not evaluated based on the MQO models since the same data that were used to calculate the limits would be used to fit the models, thereby biasing the results.

On average, the limits that were calculated using all of the data tended to be lower than those calculated using the original subset of data for both single-laboratory and interlaboratory limits. The only methods that frequently yielded higher limits when calculated using all data were Methods 200.7 and 300.0, and this generally occurred only for detection limits. The single-laboratory LCMRL generally did not follow the pattern of the other limits; LCMRLs determined using all of the data tended to be higher than the original values for Methods 335.4 and 608, and lower than the original values for Methods 200.7 and 300.0. The LCMRL differs from the other limits in that there is no extrapolation; that is, the limit can never be lower than the lowest spike level used in the calculation and can never be higher than the highest spike level used in the calculation. Therefore, this limit is more sensitive to the choice of spike level than the other limits.

Somewhat surprisingly, calculating a limit using all data did not appear to be much more likely to generate a limit than calculating a limit using the original subset of data. For example, only one more single-laboratory LCMRL was generated by using all data than was generated using the original laboratory-chosen spike levels. For some analytes, such as silver, this may be due to decreased performance at the highest spike levels, which approached the upper end of the instrument range. Additionally, the limits calculated using all data would be more likely to be affected by an outlying data point, which could cause the procedure to fail to generate a limit.

The spike levels that were originally used to calculate the single-laboratory limits were chosen by the laboratories themselves and, as a result, the levels tended to differ widely between labs for the same analyte. Therefore, the wide variability in laboratory limits for a given analyte may have been due to differences in choices of spike levels, rather than in the performance of the laboratories themselves. To assess this, the variability of laboratory limits that were determined using all of the data was compared to the variability of those determined using the original, laboratory-chosen spike levels. Pooled RSDs (calculated as the square root of the mean squared RSDs) were determined for each method and limit, and are presented in Table II d iv f in the Appendix. RSDs between the original single-laboratory limits compared to RSDs between the alternative single-laboratory limits are also presented in Figures II d iv c and II d iv d of the Appendix for the LCMRL and SLIQE20, respectively.

The RSDs of the limits that were calculated from all data tend to be slightly lower than the RSDs of the original calculated limits. The difference tends to be largest for Method 608, and smallest for Methods 300.0 and 335.4. The LCMRLs tended to vary slightly less between laboratories than other limits for most of the methods. However, the variability of limits calculated using all data was still fairly large, with pooled RSDs exceeding 70% for all limits and methods. This suggests

that much of the variability observed in the limits is due to differences in laboratory performance, rather than differences in the lab-chosen spike levels.

The Effect of Experimental Design on the ACIL Procedure

The evaluation of the ACIL procedure differed from that of the other procedures in that the limits were determined from data generated using samples prepared by the laboratory, rather than data generated using blind samples. These laboratory-spiked samples did not provide the opportunity to analyze alternate spike choices or design scenarios.

Variability between ACIL limits tended to be lower than for the other limits. The ACIL ML is heavily influenced by the choice of spike level, as the final limit is either the initially chosen spike level or an adjusted spike level chosen as a result of MQO failures. These spike level choices, and the resulting MLs, tended to vary most widely for Method 200.7, and least widely for Method 335.4. The wide variability from lab to lab for Method 200.7 is less due to the actual method variability and more related to the large inherent variability in ICP/OES instrument sensitivities. Some laboratories expressed some concerns regarding the effect of spike level choice on the resulting limits; these comments are presented and discussed in Section II d. i.

As was the case with the ACIL ML, the ACIL MDL tended to be less variable than the other detection limits. For uncensored methods, the ACIL MDL was not affected by spike level choice, because the limit was determined from blank sample results for all laboratories. The ACIL MDLs determined for Method 200.7, however, did not tend to be less variable than the ACIL MDLs determined for the censored methods (the ACIL MDLs for Method 335.4 were quite precise, however, with an RSD of 9% between laboratory ACIL MDLs). The variability between ACIL MDLs for Method 200.7 may have been due to the wide range in the number of blanks used to determine the ACIL MDLs, as this directly affects the tolerance limit multipliers used to determine the ACIL MDLs which can cause bias if the blank data follows a non-normal distribution.

Due to time constraints, the ACIL procedure had to be adapted for the Pilot Study. As a result, laboratories could not perform ongoing verification over the length of time outlined in the ACIL procedure (see Section II. C for details on limit calculation). Additionally, laboratories were not able to adjust the ML spike level if the twenty ongoing replicates failed one or MQOs. For example, some laboratories failed to meet the criterion of RSD<20% for many Method 608 and Method 625 analytes. In these cases, the estimated RSD in the confirmation analysis would be expected to exceed 20%. However, this exceedance would not truly represent the ACIL procedure; in practice, the laboratory would adjust the spike level and analyze replicates at the new level upon such a failure.

Effect of Outlier Removal

The effect of outlier removal was assessed by performing all limit calculation and confirmation analyses with and without outlier removal. While the study MQOs were somewhat more likely to be met if outliers were removed, the effect was minimal. In many cases, the MQO performance was worse (i.e., the MQO statistic was further from the study criterion) when outliers were removed. The reason for this tendency is that while outlier removal will tend to lower the variability of the data and remove possible false positive and false negative results, the outlier removal may also decrease the calculated limits (or make it possible that a limit can be calculated). Generally, outlier removal had the largest effect on the false positive and false negative rates, because these MQO statistics are more heavily influenced by the tails of the data distribution than the mean and RSD. Outlier removal also had a slightly larger effect on the

single-laboratory limits. This difference was likely due to the larger number of individual values used to calculate and confirm the interlaboratory limits than the single-laboratory limits and, therefore, a single outlying result would have a greater effect on a single-laboratory limit.

v. Does the Procedure achieve its intended purpose? (ZE/KM)

ACIL Procedure

The ACIL procedure was developed to establish an intra-laboratory (single laboratory) detection and quantitation limit. The procedure is designed to determine a detection limit (MDL) with a false positive rate of 1%. The procedure utilizes method blanks for the determination of Lc when available. Also, the ACIL procedure determines precision and accuracy at the quantitation limit. Lastly, it utilizes long term data and addresses the case of non-zero blanks.

The ACIL procedure was revised to include a procedure for estimating the Quantitation Limit (QL); similarly to the ACIL MDL, the ML used the same MQOs as the pilot. Therefore, the evaluation of the procedure versus pilot study MQOs in section II.d.viii is identical to the evaluation of whether or not the procedure achieved its intended purpose.

ASTM IDE (Single Lab Procedure)

The ASTM IDE procedure was designed to provide a high probability that results of the analytical methods studied will produce values that exceed the interlaboratory detection estimate that represent the presence of an analyte in the sample (approximately 99%). The procedure is designed to produce two detection limits (the Yc and Lc, where Yc is the measured concentration detection limit and Lc is the true concentration detection limit) with a false positive rate of 1%. Also, it was derived to yield a true concentration (the IDE) at which there would be a 5% false negative rate when making the detection decision at the Lc. For the purpose of the pilot study, the procedure was adapted for a single laboratory detection/quantitation limit.

The intended purpose of the Lc and Yc match the MQO for false positive rate, and therefore the evaluation of these limits matches the one in the MQO section. The difference between the Lc and Yc is that the Yc represents the critical level at the measured concentration and the Lc represents the critical level at the true concentration (corrected for bias). The IDE, however, was designed to achieve a false negative rate of 5%, rather than the target 1% rate MQO chosen for the study. While the false negative rates at the IDE frequently exceeded the 1% MQO, they tended to fall below 5% the majority of the time for most analytes and labs. The mean modeled false negative rate (based on Yc) was approximately 5% or below for four of the five methods if outliers were kept, and was approximately 5% or below for three of the five methods if outliers were removed. The false negative rates based on Lc tended to be higher, with mean modeled false negative rates exceeding 5% for most methods, regardless of outlier removal.

Since the target FN error rate was 5%, if the procedure met its objectives we would anticipate the average and median false negative error rate to be close to 5%, with the individual FN error rates falling above 5% about half of the time and below 5% about half the time. According to the summary table below showing false negative error rates for the single laboratory estimates using the ASTM procedure, the procedure failed to achieve its designed objective of 5% FN the majority of the time, for all methods and analytes. The best performance was with method 200.7 where 36% of the time the FN rate was in the 1% to 10% range when using modeling to estimate the FN error rate.

See ASTM False Negative Error Rate at Lc for Results at IDE Single Laboratory Estimates – Outliers Removed in Appendices.

ASTM IQE 20%, 30% (Single Lab Procedure)

The ASTM IQE procedure was developed to estimate the concentrations at which the Relative Standard Deviation (RSD) would be 20% and 30%. The procedure uses multiple linear and nonlinear models for standard deviation and a weighted linear regression model for mean recovery.

For most methods, the mean modeled RSD at the IQE20 was below 20%, and the mean modeled RSD at the IQE30 was below 30%. The mean modeled RSDs at the IQEs were much lower than the target values for Method 300.0, indicating that the limits were overestimates of the minimum concentration to yield the target RSD. The mean modeled RSDs at the IQEs exceeded the target values for Method 335.4 if outliers were removed, indicating that the limits were underestimates of the minimum concentration to yield the target RSD.

Single lab LCMRL

This procedure was designed to obtain the lowest true concentration for which future analyte recovery is predicted to be in the range of 50-150% with 99% confidence. This differs from the recovery MQO because it targets individual recovery rather than mean recovery.

The calculated lab LCMRLs were evaluated for target performance by determining the percent of Pilot Study results spiked at levels exceeding the LCMRL for which the recovery exceeded 150% or fell below 50%. These percents were above the target 1% for Methods 608 (1.17%) and 335.4 (2.42%) only. When outlier removal was performed, the percents were below 1% for all methods.

The LCMRL procedure did not produce a limits for analytes and for which recovery at low concentrations never approached 100 %. This occurred most frequently for Methods608 and 625, and primarily for Interlaboratory data as the LCMRL was designed to function primarily as a inter laboratory procedure.

Single Lab Hubaux-Vos

The Hubaux-Vos detection limit procedure is designed to use a graph to determine two sensitivity limits: Yc and Ld. The graph is composed of plotting measured versus true concentration. The false positive and false negative rates are predicted at 0.5% using the scatter plot from the LCMRL.

During the Pilot Study, only the Yc, and not the Ld, was calculated. For the majority of the methods, the median false positive rate for the Yc was less than 0.5%; indicating the limit met or slightly exceeded the target value. For Method 200.7, the median false positive rate was greater than 0.5%, with and without outlier removal. For all methods, the mean false positive rate exceeded 0.5%, indicating that there were a few analytes and labs for which the false positive rate was much higher than the target value.

IL- LCMRL

The interlaboratory LCMRL targets the same 50-150% recovery range for individual sample recovery, but for all laboratories. This limit could not be calculated for the majority of analytes for Methods 608 and 625, due to the lower recoveries for these methods, and the large variability between laboratories. Where interlaboratory LCMRLs were determined, the percentage of results at or above the limit for which recovery fell outside the 50-150% window exceeded 1% for Methods 608 and 625 without outlier removal, and for Method 625 only with outlier removal.

IL- ASTM IDE/IQE20, IQE30 Interlaboratory Procedures

The procedure was designed to provide specific estimates for the interlaboratory critical level (Yc and Lc), detection level (IDE) and three estimates of quantitation (IQE_{10} IQE_{20} IQE_{30}). The procedure specifically targets a false positive rate at Lc and Yc, false negative error rate at (IDE) and precision at the IQE across the entire range of an analytical method.

The intended purpose of the Lc and Yc match the MQO for false positive rate, and therefore the evaluation of these limits matches the one in the MQO section. For the IDE, there were extreme mixed results observed for all methods and analytes. There were times that the target 5% false negative rate was achieved (EPA Methods 300.0, 200.7, and Method 625 based on Yc), and times that they were not achieved (EPA Methods 335.4, 608 and Method 608 based on Lc).

According to the summary table below showing false negative error rates for the interlaboratory estimates using the ASTM procedure, although performance was better than the single laboratory estimate, the procedure failed to achieve its designed objective of 5% FN the majority of the time, for most methods and analytes. The best performance was with methods 335.4 (modeled), 200.7 and 608.

See ASTM False Negative Error Rate at Lc for Results at IDE Interlaboratory Estimates - Outliers Removed in Appendices.

The ASTM IQE_{20} IQE_{30} procedure had mixed results. Similarly, to the IDE, the IQEs tended to achieve the target RSDs for Methods 300.0 and 200.7 and tended to not achieve the target RSDs for Methods 335.4 and 608. The ASTM procedure was unable to obtain an IQE_{20} most frequently for Method 608, with no limit being determined for approximately 40% of the analytes. The ASTM procedure was unable to obtain an IQE_{30} for 5-20% of analytes in Methods 608 and 625. The inability of the ASTM procedure to obtain an IQE estimate was due to the performance of the analytical method, because 20% or even 30% RSD was never achievable at any concentration tested for several analytes.

vi. Does the procedure work for all different types of analytical methods? (RB)

Summary

The ACIL procedure when used correctly, performed well for all methods. In part this is because the ACIL procedure was expressly designed to meet the MQOs identified for the pilot (although it could be readily modified to meet other MQOs). In part it was due to avoiding the need to extrapolate and because the uncensored procedure takes blank bias directly into account. In many cases failures to meet the MQOs when using the ACIL procedure were due to failure to follow the procedure correctly, especially the direction that the QL must be at least 2 times the DL. Other failures of the ACIL and other procedures could be due to intermittent blank contamination problems, and indicate the need to incorporate intermittent blank guidance into whatever procedure is finally recommended to the FAC and EPA. The LCMRL performed well for quantitation limits. There were quite a few instances for which the LCMRL could not be calculated: these could be considered failures of the analytical method rather than the LCMRL, but still there needs to be some way of dealing with poor performing analytes in current methods. The Hubbaux Vos procedure paired with the LCMRL for detection limits performed less well, high rates of both false positives and false negatives were observed. The ASTM IDE and IQE procedures did not fare well when applied on a single lab basis. It is recommended that the pooled multi-lab estimates be compared to ASTM and LCMRL inter-laboratory estimates, using all samples at all spike levels to assess the applicability of these procedures. This will also allow the

evaluation of the robustness of the procedures when there is a wide discrepancy between the single lab detection and quantitation limit values.

In general, the ACIL procedure, with addition of intermittent blank guidance, should perform well for all methods. The LCMRL performs well for setting the quantitation limit, assuming that the analytical method can meet the level of performance required. The Hubbaux Vos procedure can be effective if the precision of blank measurements is as predicted by spike measurements, but fails if that is not the case. It also fails if the intercept is not close to the actual blank bias level. The ASTM procedure was not designed to be applied to the single lab case, and in general did not perform well for this function.

The pilot study results indicate that any procedure should include the following requirements:

- The QL should be greater than the DL by some factor, at least 2-3 times higher and may be recovery dependent.
- Routine blank results should be checked against the calculated DL
- Techniques to accommodate intermittent blank contamination problems need to be incorporated
- The DL and QL should be checked against ongoing data periodically.

Details

The primary focus of this discussion will be on the analytical methods tested in the pilot study. From this analysis we will extrapolate to other methods not included in the pilot study.
First, we need to define what we mean by "work". For the purposes of this discussion a procedure is deemed to work if the MQOs of the pilot study are met. In some cases, the procedures actually target different MQOs from those defined for the pilot study. These variations will be examined in the discussion for each procedure. The MQOs for the pilot study were as follows:

False positive rate

Less than 1%. A false positive is defined as an unspiked reagent water blank processed through the analytical method that gives a result above the detection limit determined by the procedure.

False Negative rate

Less than 1%. A false negative is defined as a sample spiked at a concentration at or above the quantitation limit determined by the procedure that gives a result less than the detection limit.

Precision

Less than 20% RSD. Defined as the precision of replicate measurements for spiked samples at the quantitation limit

Accuracy

Between than 50-150% of true value. Defined as the mean recovery for replicate measurements of spiked samples at the quantitation limit

Analytical Methods

The analytical methods chosen for the pilot study were intended to illustrate a wide range of detection and quantitation issues.

Method 200.7

This is an example of a no-censored method, i.e., numerical results (which can be negative) are always obtained for method blanks. In addition, some analytes in this method have levels of

interest that are very close to the ultimate sensitivity of the instrumentation (As, Se, Pb, Tl, etc), while others are particularly subject to blank contamination (Cu, Zn, Fe, etc).

Method 300.0
This is a chromatographic method that often shows very good precision in the short term. Many laboratories observe that the 40CFR Part 136 MDLs obtained for this method are well below the level at which a chromatographic peak can be determined. As configured in environmental laboratories, the method is usually censored (ND results obtained for blanks) but the ubiquitous nature of some of the analytes (e.g., chloride) may result in very frequent detection for blanks.

Method 335.4
This single analyte method has been observed to be particularly prone to false positives in routine laboratory operations.

Method 608
A chromatographic method with a selective detector that can be subject to varying levels of noise dependent on the cleanliness of the detector and which can be configured as either censored or non-censored.

Method 625
A GC/MS method with compound identification criteria (qualifier ion co-elution, erc) that usually causes the method to be of the censored type.

Method 200.7 details

False positives
Overall, the ACIL procedure performed better (less false positive rates above 1%) than the Hubbaux Vos or ASTM procedures. This is not surprising given that the ACIL detection limit calculation is based directly on the variability of the blanks, while the HV and ASTM procedures are based on modeling from spiked samples but could be extrapolate below the lowest spike for censored methods. The few occasions where the ACIL false positive rate was considerably above 1% objective were mostly due to calibration blanks being included in the assessment as well as method blanks. This discrepancy is under further investigation. Both the HV and ASTM procedures can result in high false positive rates if the intercept is significantly different from the actual blank bias. Most of the high false positive rates for the HV and ASTM procedures were observed for analytes where positive blank bias might be expected (Ca, Na, K, Al, Zn). This leads to some concern for the reliability of these procedures for tests in general where positive blank bias is a significant driver of the true detection limit. There types of methods are becoming more prevalent as required levels of detection are driven lower. For example, ICP/MS detection limits for the same instrument may vary by 2 or 3 orders of magnitude between use with sub-boiling distilled acids in a clean room environment and use of reagent grade acids in a typical environmental lab environment. PCBs by method 1668, mercury by 1631 and dioxins by 1613 are other examples of methods where detection limits are likely to be highly dependent on blank bias.

False negatives
The ACIL procedure performed very well (few false negative rates > 1%) and the LCMRL also performed well in general with a slightly higher occurrence of high false negative rates. In a few cases, the HV detection limit was higher than the LCMRL, which results in a false negative rate of 100%. This is clearly a serious problem, but probably more due to difficulties with the HV procedure than any problem with the LCMRL. In a few occurrences of high false negative rates

for the ACIL procedure, the ACIL quantitation limit was less than 2 times the ACIL detection limit. This is an incorrect application of the ACIL procedure, which requires that the QL be at least 2X the DL. If the QL has been properly elevated in these cases, the false negative rate would have been much lower. The ASTM procedures generated much higher false negative rates, in general because there is no control on the QL being close to or even below the DL, although a study supervisor should never accept an IQE, which is less than and IDE since a quantifiable value by definition must be detectable. On the single lab basis, the ASTM procedure did not perform well for the 1% FN MQO although it should be noted that the procedure targets a 5% FN error rate at the IDE and not at the IQE.

Precision

The precision MQO was met for almost all analytes by both the ACIL and LCMRL procedures. About one third of the analytes exceeded 20% RSD for the ASTM IQE20. Since this procedure targets 20% RSD, it is not surprising that many analytes exceed the limit. However some of the analytes exceeded the 20% RSD target by a large margin, up to over 200% RSD.

Accuracy

Almost all analytes met the 50-150% mean recovery MQO for both the ACIL and LCMRL procedures. About 15% of analytes failed the criteria for the ASTM IQE20, because of extreme low or high recovery at the targeted %RSD. It should be noted that the ASTM procedure does not target any particular recovery MQO, so a greater frequency of failures is not surprising.

200.7 Summary

The ACIL and LCMRL procedures worked effectively for the precision and accuracy MQOs and should be expected to work effectively for methods with similar characteristics, namely typically good accuracy and precision across the quantitative range. This includes most methods for metals analysis and most methods that do not have separate preparation steps that introduce a substantial amount of variability such as some analytical methods for organics.

The ACIL procedure was most effective at meeting the false positive criterion because of the basis on blank data for uncensored methods, rather than extrapolation from spikes. In addition the on-going portion of the ACIL procedure, which was not tested basis the false positive estimate on method blank data for all methods. The ACIL procedure was also fairly effective at meeting the false negative criterion, in part due to the requirement that the QL be at least 2X the DL. Other procedures would have been more successful at meeting this MQO if a similar requirement was included. The ASTM procedures were less effective for this analyte set for single lab data.

Method 300.0

False positives

Several analytes had a high false positive rate by the ACIL procedure because of a high mean blank value that was not compensated for using the censored procedure. Incorporation of the intermittent blank guidance (or something similar) into the procedure should effectively address these issues. A similar frequency of high false positive rates was noted for the other procedures (LCMRL and ASTM). This is in part due to the intermittent nature of the blank contamination (ie blank variability greater than would be expected based on the spike sample results) and in part due to intercepts that are different from the mean blank values. The intermittent blank procedure would also have to be incorporated into these procedures in order to eliminate high false positive rates due to this mechanism.

False negatives
The false negative MQO of <1% was met effectively by both the ACIL and LCMRL procedures. The ASTM procedure produced a large number of analytes with high false positive rates, generally because the IDE was too close to the IQE.

Precision
The pilot study MQO precision goal of <20% RSD was met in general by both the ACIL and LCMRL procedures. Most analytes also met the goal using the ASTM IQE 20 procedure, although some of the failures had very high RSD (>100%).

Accuracy
The recovery goal MQO was met most of the time by both the ACIL and LCMRL procedures. About one third of analytes failed the criterion using the IQE 20 procedure, some by a considerable margin because of extreme low or high recovery at the targeted %RSD.

Method 300 summary
Intermittent blank detections caused problems for all procedures and illustrated the need to incorporate techniques for dealing with the issue into the final detection limit procedure. The quantitation MQOs for the pilot study were mostly met using both ACIL and LCMRL procedures, and other chromatographic techniques without separate preparation steps should be expected to perform similarly.

Method 335.4

False positives
Most labs obtained a low false positive rate using all procedures. One lab had a high false positive rate for HV and ASTM Lc procedures due to underestimation of blank variability based on the spiked data.

False negatives
Several labs obtained high false negative rates using the ACIL procedure. In some cases this was due to not following the procedure correctly (did not ensure that the QL was at least 2X the DL). In another case it was due to the ongoing spikes having RSD > 20%, which in practice would lead to an increase in the QL when following the ACIL procedure.
There were also some high false negative rates obtained using the LCMRL procedure. In some cases this appeared to be due to outliers, in one case due to a HV detection limit that was greater than the LCMRL QL.
The ASTM procedure was also impacted by outliers, and in some cases the QL was less than 2X the DL.

Precision
The LCMRL more often obtained the 20% RSD goal than the ACIL procedure. However, in two cases the LCMRL could not be calculated, probably due to the same high variability that caused the ACIL precision failure. In both cases, the procedures would call for increasing the QL for ongoing data to a level at which the precision goal was met. The IQE 20 performed well for this method, with a precision < 20%RSD in all cases except one.

Accuracy
The accuracy goal was most met for this method using all procedures.

Method 335.4 summary.
This method illustrated the need to reassess detection and quantitation limits once a larger quantity of data is collected. In many cases the DLs and QLs for the ACIL procedure would have changed based on the larger data set. Assuming that this reassessment is a requirement, then the procedures should work effectively for methods of this type.

Method 608

False positives
Some analytes at one laboratory has a high false positive rate because the censored ACIL procedure was used even though the blank data were uncensored. This is an incorrect use of the ACIL procedure. When correctly used to calculate a DL based on the uncensored ACIL procedure the false positive rate becomes low. Most of these analytes had an even higher false positive rate using the HV procedure. The ASTM and HV procedures also produced high false positive rates in some cases due to non-zero intercepts.

False negatives
In a few cases the ACIL procedure generated high false negative results. These were mostly due to failure to follow the direction in the procedure that the QL must be at least 2X the DL. The LCMRL and HV procedures produced a somewhat higher incidence of high false negative rates, mostly due to the QL calculating too close to the DL.

Precision
Endosulfan I and Endosulfan II had high RSDs in general. This resulted in failure to meet the RSD MQO using the ACIL procedure and inability to calculate a LCMRL. Even excluding Endosulfan, several analytes at some laboratories failed to meet the pilot study precision MQO using the ACIL procedure. (Note that using the ongoing part of the ACIL procedure, this would have resulted in a requirement to raise QLs for these analytes). For the same high RSD reasons, a LCMRL could not be calculated for some analytes at some laboratories. The ASTM procedure failed to meet the precision MQO in many cases due to the same problem – essentially poor precision at any concentration.

Accuracy
The mean recovery MQO was met in almost all cases using the ACIL procedure. This was also the case for the LCMRL (noting that there were several cases for which the LCMRL could not be calculated due to the poor precision exhibited by the analytical method. Most analytes also met the recovery MQO using the ASTM procedure, although a few failed by a wide margin because of extreme low or high recovery at the targeted %RSD.

Method 608 summary
This data indicated that while the pilot study MQOs of 50-150% mean recovery and 20% RSD may be considered quite liberal , however they were not achieved by several laboratories for some 608 analytes. It may be necessary to set even wider MQOs for these methods, or alternatively call the methods semi-quantitative. Rather than considering that the detection/quantitation procedures failed, it should be considered that the analytical methods failed to provide the quality of data that was specified for quantitation.

Method 625

False positives
Since method 625 is generally censored, the false positive rates were low for all procedures. A few exceptions were due to intermittent blank contamination issues(phthalates).

False negatives
Several analytes produced high false negative rates using the ACIL procedure. In most cases this was because the QL spike level has mean recovery and/or RSD that failed the MQOs, or a QL that was less than 2 times the DL. In practice this would cause the ACIL QL to be raised, thereby bringing the false negative rate under control, but there was insufficient time in the pilot study for this part of the procedure. The LCMRL had a much lower false negative rate, but in general this appeared to be due to the fact that no LCMRL was calculated for the poorer performing analytes. Many analytes had high false negative rates following the ASTM procedures, usually this was because the IQE20 or IQE30 was too close to the IDE. The ASTM procedure targets a 5% FN error rate at the L_C for measurements at the IDE as defined by Currie. If the IQE10 were evaluated the false negative error rate would have been much lower.

Precision
Most analytes met the precision MQO following the ACIL procedure. Benzidines, phenols and phthalates were the most common failures. Poor precision resulted in failure to calculate a LCMRL in many cases. The IQE20 procedure usually met the precision MQO.

Accuracy
Most analytes met the accuracy MQO using the ACIL procedure. Failures were usually benzidines, phenols and phthalates. Analytes with a calculated LCMRL met the accuracy criteria, but many analytes did not obtain a calculated LCMRL. Most analytes met the accuracy MQO using the ASTL IQE 20 procedure.

Method 625 summary
This method includes several analytes that exhibit poor precision and accuracy across the analytical range. Some allowance needs to be made for these analytes in existing methods, either wide MQOs for precision and accuracy, or improvements in methods must be made or acceptance that some data produced will be semi-quantitative. One example of making method improvements would be to require continuous liquid/liquid extraction for the acid fraction (PSR.I.d.vi.Phenol Analysis by Method 625). The quantitation limit procedure needs to include direction on how to identify these analytes, and how to appropriately handle and communicate the quality of data obtained.

vii. Does the procedure work if applied to real world sample matrices? (LL)

Because of budget limitations, none of the procedures were evaluated using real world matrices and therefore, the pilot study does not provide any information on the applicability of the procedures if applied to real world matrices.

viii. MQO's (JPH/KM/JPL)

1. Did the procedure meet the bias at L_Q established by the FACDQ?

The MQO for bias at L_Q for the pilot study was deferred to the Technical Workgroup (TWG) by the FACDQ. The TWG established the bias MQO at 50-150% mean recovery.

1.1. What is the data? What Works? What doesn't work? Confidence levels?

The pilot study results as related to the bias MQO are summarized graphically in the following attachments.

For Interlaboratory bias;

PSR.II.d.viii.1.a.(Interlaboratory Bias methods 300.0 and 335.4)
PSR.II.d.viii.1.a.(Interlaboratory Bias method 200.7) PSR.II.d.viii.1.a.(Interlaboratory Bias method 608) PSR.II.d.viii.1.a.(Interlaboratory Bias method 625)

For Single Laboratory bias;

PSR.II.d.viii.1.a.(Mean Laboratory Bias methods 300.0 and 335.4) PSR.II.d.viii.1.a.(Mean Laboratory Bias method 200.7) PSR.II.d.viii.1.a.(Mean Laboratory Bias method 608) PSR.II.d.viii.1.a.(Mean Laboratory Bias method 625)

For MMA PCB Study Bias;

PSR.II.d.viii.1.a.(Interlaboratory and Mean Laboratory Bias method 608)

For each method, interlaboratory bias is depicted in four graphs, based on the following breakdown:;

- With Outliers (all data; no statistical outliers removed) with bias estimated by modeling (Modeled)
- With Outliers, with bias estimated by interpolation (Interpolated)
- Outliers Removed (statistical outliers removed from data) with bias estimated by modeling (Modeled)
- Outliers Removed, with bias estimated by interpolation (Interpolated)

Each graph depicts the percent recovery (identified by a diamond) for each analyte of the method(s) specified by the title of the attachment. The X-axis lists analyte, grouped by each of the interlaboratory quantitation procedures evaluated. When a small diamond is either off scale or located on the zero percent recovery axis it indicates that the procedure was unable to yield a valid quantification limit estimate. Ideally if the procedure met the FACDQ pilot study MQOs for bias, all diamonds would fall between 50 and 150% percent recovery. A diamond on the 100% line on the graph indicates that there was no bias at the given limit for that analyte.

The Mean Laboratory Bias graphs depicting single laboratory bias are formatted similarly to the interlaboratory bias graphs. Four graphs are also presented for each analytical method(s) evaluated. However instead of small diamonds the mean laboratory bias for a given analyte is identified by a small dash. For each analyte, error bars extend from the mean plus or minus one standard deviation, where standard deviation is calculated from the estimated recoveries at each of the laboratory limits. Similarly to the interlaboratory graphs, a dash on the 100% line would indicate that on average, there was no bias at the calculated laboratory limits. No error bars would either indicate that the bias at the lab limits was the same for all laboratories, or that only one laboratory limit could be calculated for the given analyte.

The Michigan Manufacturing Association (MMA) Polychlorinated Biphenyl (PCB) Study data by EPA method 608 are presented using a similar chart design as those described above. However, the Y-Axis depicts data with outlier removal (OR), without outlier removal (WO), modeled (m) and interpolated (i). The average laboratory bias chart includes both PCB Aroclor 1016 and 1260, but the interlaboratory bias charts present each of the PCB Aroclors separately. The same rules apply when interpreting the results.

Data Analysis Findings

Outlier Removal for Interlaboratory Procedures – Across all methods and analytes, outlier removal had minimal or no impact on how well the quantitation limit procedures performed in meeting the bias MQO (see discussion in Section II d. iv). The exceptions to this rule include;

- Total cyanide by method 335.4 (modeled), outlier removal allowed a valid LCMRL to be calculated.
- Chloride by method 300.0 (modeled), outlier removal caused the IQE30 to have more than 150% bias.
- 4.4'-DDE by method 608, outlier removal allowed a valid IQE20 to be calculated.
- Endrin by method 608, outlier removal allowed a valid IQE20 to be calculated.
- Di-N-Butylphthalate by method 625, outlier removal allowed a valid LCMRL to be calculated.

Outlier Removal for Single Laboratory Procedures – Across all methods and analytes outlier removal had minimal or no impact on how well the quantitation limit procedures performed in meeting the bias MQO. The exceptions to this rule include;

- All method 300.0 parameters (interpolated), outlier removal significantly improved the ability of the IQE20 to achieve bias objectives.
- Aluminum (modeled) and Copper by method 200.7, outlier removal allowed acceptable recoveries to be achieved and the bias MQO to be achieved. This was because one lab had extremely low recovery for Aluminum (-2824%) and another lab had very high recovery for Copper (567%), heavily influenced by a single outlying result.

Outlier Removal for MMA PCB Method 608 Data – Mean single laboratory bias and variability for IQE20 and IQE30 using interpolation was improved with outlier removal. This appears to have been due mainly to a single laboratory with a strongly high-biased result (525% recovery for Aroclor 1016) prior to outlier removal. The mean RSD at the IQE20 and IQE30 exceeded 20% and 30% respectively prior to outlier removal, but fell within the target RSD after outlier removal. An interlaboratory IQE20 limit could not be calculated after the removal of outlying results for PCB Aroclor 1260. A valid IQE20 that met the bias MQO could be achieved for Aroclor 1260 when both outlying labs and results were removed.

Modeling versus Interpolation – In general, there was minimal observable effect whether the modeling or interpolation techniques were used to evaluate performance of the procedures. The only exceptions to this rule are as follows:

- Many method 300.0 parameters showed less single lab bias variability when modeling was used in evaluating both the IQE20 and IQE30.
- Copper by method 200.7 (with outliers), showed substantially less single lab bias and bias variability when modeling was used for the IQE20. The pilot MQOs were achieved for the IQE20 with modeling, but not by interpolation for this parameter.

Generally, the difference between modeled and interpolated recovery estimates will be small unless there are large increases or drops in recovery between two consecutive spike levels. This happened most frequently for Method 300.0, which often displayed large increases in recovery followed by a sharp decrease. For this method, the interpolated recovery estimates are likely more reliable.

General Observations

Interlaboratory Bias – The LCMRL and IQE20 consistently achieve the FACDQ MQO of 50-150% recovery for all analytes and methods. The IQE30 met the bias objectives the majority of the time. The mean recoveries at the different quantitation limits tended to be close to 100% for most analytes and labs. Method 300.0 tended to have slightly high bias overall, with chloride bias approaching 150%. The majority of labs had a high bias at low spike levels for this analyte, such that the interlaboratory variability (as expressed by RSD) was low, in part because the results tended to be high biased. Both method 608 and 625 had low bias, with mean recoveries at quantitation limits for method 608 analytes ranging from 75% to105% and mean recoveries at quantitation limits for method 625 analytes ranging from about 50% to 90%.

The LCMRL could not be calculated for many analytes in methods 608 and 625. In method 608, 15 out of 18 analytes could not achieve 50% recovery with high probability based on interlaboratory data and in method 625 17 out of 18 analytes could not achieve 50% recovery with high probability based on interlaboratory data. This was due to some of the laboratories having low recoveries and poor precision throughout the concentration range for these methods.

The IQE20 also was unable to obtain a valid result for 8 of the 18 method 608 analytes) and 3 of the 18 method 625 analytes (2,4-Dinitrophenol, 3, 3'-Dichlorobenzidine and Phenol). Interlaboratory bias MQOs were never (or rarely) achieved for Alpha-BHC, Endosulfan I, Endosulfan II and Heptachlor Epoxide by method 608. Valid limits could not be calculated for these analytes because of the large interlaboratory variability between labs through the concentration range. For example see the attached phenol data by method 625, which demonstrates the inability for the method as performed by the pilot study labs to achieve 20% RSD at any concentration evaluated (*PSR.II.d.viii.1.a.(IL %RSD vs Conc Method 625 Phenol)).* It should be emphasized that unlike the OGWDW LCMRL, the ASTM IQE does not have a bias objective, although this could be added to the procedure.

Single Laboratory Bias – The average laboratory bias MQO of 50% to 150% recovery was met for nearly all procedures (ACIL-ML, LCMRL, IQE20 and IQE30), methods and analytes with exceptions noted below.

- Mean bias for most method 300.0 analytes exceeded 150% for the IQE20 and IQE30 when evaluated using interpolation, and mean bias for 30-50% of analytes and labs exceeded 150% for the IQE20 and IQE30 when evaluated using modeling.
- Even after outlier removal average calcium and zinc recoveries (method 200.7) were greater than 150% for the IQE20 and IQE30.
- All method 625 analytes had average recoveries between 50% and 100%, other than 2,4-Dinitrophenol, 3, 3'-Dichlorobenzidine and Phenol for which mean recoveries were near 40%.
- For method 625 the LCMRL tended to yield percent recoveries closer to 100% and the IQE30 tended to yield percent recoveries closer to 50%.

For the Method 300.0 and 200.7 analytes cited above, the laboratories often yielded fairly precise results despite the high bias, such that the precision criteria for these limits were met. In addition, the recovery correction often decreased the quantitation limits for these analytes, because the recovery often exceeded 100% at the estimated limits.

MMA PCB Study Method 608 - Overall the LCMRL, IQE20 and IQE30 achieved pilot study bias MQO for both single and interlaboratory procedures.

1.1.A. If it fails; why?

Interlaboratory Procedures

OGWDW LCMRL

When the LCMRL was unable to meet the pilot MQO for bias or calculate a valid limit it was primarily due to the poor performance of the method. However reason may be because the original LCMRL document did not include guidance to bracket LCMRL value with a spiking level. Not properly bracketing the LCMRL caused two types of errors.

1. When data does not include a low enough spike, the calculator defaults to the lowest spike level. This can be resolved by including a lower spike level in the data set. *PSR.II.d.viii.1.a.(Not Low Enough)*

2. When the spike levels used have high variance and/or poor recovery a calculator error message, "Could not determine LCMRL". Occasionally this may be resolved by including a higher concentration spike. If the laboratory was unable to meet the MQO for that analyte at any concentration, remedial action to improve data quality was needed. *PSR.II.d.viii.1.a.(High Variance)*

These shortcomings could have been resolved by the interlaboratory study design team selecting more appropriate spike concentrations for each analyte/method combination. Proper bracketing of the LCMRL the procedure still might not have been able to generate a valid limit, because when using all concentration form the study even fewer limits were calculated. This indicates that the problem is due to the low-biased recoveries and large interlaboratory variability especially for methods 608 and 625.
PSR.II.d.viii.1.a.(No More Data Available))

ASTM IQE

Although the ASTM procedure does not directly target bias as one of the performance criteria it must achieve to calculate valid limits (although high-biased results will decrease the RSD and low-biased results will increase the RSD), the bias MQO was achieved frequently in the pilot study data. On the three occasions when the IQE 30 was unable to calculate a limit or meet the pilot MQOs for bias it was because the performance of the labs for that analyte/method was not adequate to achieve the precision criteria of the procedure of 30% RSD. When all data for a given method were used, a valid IQE30 was achieved for all except two analyte/method combinations.

PSR.II.d.viii.1.a.i.(%RSD vs. Conc of 3,3'-Dichlorobenzidine by 625)
PSR.II.d.viii.1.a.i.(%RSD vs. Conc of Phenol by 625)

When all data for a given method were used, a valid IQE20 was achieved for all except seven analyte/method combinations, five Method 608 analytes and two Method 625 analytes. The failures for these analytes was due to the fact that at least 20% RSD was never achieved at any of the concentrations evaluated.

The ASTM IQE always achieved the pilot study MQO for bias with few exceptions (IQE30 for Chloride by Method 300.0, IQE30 for Benzo(a)pyrene and pentachlorophenol by Method 625),

where mean recovery between 50% and 150% could not be achieved. The cause of the chloride limit failure was two fold, first a low enough spike concentration was not selected to obtain an accurate IQE30. An IQE30 of ~200 ug/L was extrapolated when using a low level spike of 1000 ug/L. If lower level spikes were used then an actual IQE30 of ~600 ug/L would have been calculated. Second, there was very high bias at the low end of the concentration range for Chloride (below 500 ug/L), yielding percent recoveries of over 150%.

PSR.II.d.viii.1.a.i.(IQE30 for Chloride by 300 without Low Level Spike)
PSR.II.d.viii.1.a.i.(IQE for Chloride by 300 with Low Level Spike)
PSR.II.d.viii.1.a.i.(Chloride by 300 Low Level Bias)

Single Laboratory Procedures

ACIL-ML

Although an ACIL-ML was always calculated although it did not always achieve its bias objectives. This was a shortcoming of the pilot study itself, since it did not allow enough time for spike concentrations to be adjusted and reiterated. Bias objective failure rates are listed below for the ACIL procedure.

Method	<50% Recv	>150% Recv
300.0	0%	0%
200.7	0%	0%
335.4	0%	0%
608	1%	0%
625	17%	1%

This procedure was written specifically to achieve the pilot study MQOs, and it did perform remarkably well in achieving the bias MQOs of 50% to 150% recovery, when evaluated against the ongoing verification samples. In actual practice the laboratory would adjust the ML and perform ongoing verification at the new concentration. Analyte/Method combinations where the ACIL-ML was unable or struggled to achieve one of the pilot MQOs for bias include the following;

- Potassium by Method 300, while the average recovery was always well within the objectives range, individual lab recoveries ranged from near 50% to over 200%.
- Like all of the other procedures tested, 2,4-Dinitrophenol, 3,3'-Dichlorobenzidine and Phenol by Method 625 all failed with low bias.

We would have to look at the raw data to confirm this, but the reason that individual lab recoveries for potassium vary from low to high values is most likely error introduced by the calibration. Many labs use an unweighted linear regression for method 300.0, and this type of curve fit tends to introduce large errors at the low end, which can be in either direction, depending on the specific calibration curve. Recalculation of this data using an average calibration factor curve would probably reduce the bias considerably.
2,4-dinitrophenol, 3,3'-dichlorobenzidine and phenol all have low recovery across the whole analytical range, especially if separator funnel extraction is used. Therefore the MQOs would not be met at any concentration.

OGWDW LCMRL

The high cases where an LCMRL could not be calculated in methods 608 and 625 is primarily due to the poor performance of these methods. It may also be explained in part by the same two

procedural short comings associated with bracketing the LCMRL with spikes, as identified for the interlaboratory data. For the single laboratory determinations this could have been resolved by providing additional guidance to the labs in the procedure or as part of the pilot study design. The OGWDW LCMRL targets 50% to 150% recovery for all laboratories, therefore it should easily achieve the pilot study objectives for average bias. The procedure did perform very well when an LCMRL was determined, with the exception of the following Analyte/Method combinations.

- Calcium by Method 200.7, had one lab out of eight exceed the upper bias limit for three out of four datasets. This may be the result of an extrapolation error.
- Aldrin by Method 608 yielded a high bias for one out of three laboratories.

ASTM IQE

The IQE 30 was unable to calculate a valid limit on only five occasions;
Lab 35 for Endosulfan I by Method 608
Labs 29, 31 and 35 for Endosulfan II by Method 608
Lab 39 for Phenol by Method 625
However, when all study data for a given Analyte/Method combination was used only a single laboratory failed for a single parameter (Lab 29 for Endosulfan II by Method 608).

PSR.II.d.viii.1.a.i.(%RSD vs. Conc of Endosulfan II by Lab 29)

The IQE20 was unable to calculate a valid limit more often, due to its more stringent precision requirement. The IQE20 was unable to calculate a limit for seven labs using Method 608 and six labs using method 625.

When the IQE20 or IQE30 failed to meet the pilot MQO for bias it was caused by two factors;

1. Poor performance of the laboratories for selected analyte/method combinations.
2. Due to limitations in the pilot study design related to spike concentration.

For some Method 300.0 analytes high bias occurred at low concentrations and for some Method 625 analytes low bias occurred at high concentrations. However, the bias MQO was generally achieved at a lower concentration in the data than the precision or false negative rate MQOs. The spike concentrations for determining the IQE were derived from the spikes chosen by the lab to achieve the LCMRL, and therefore may not have been optimized. As a result of these factors the bias MQOs of 50% to 150% recovery were generally, but not always achieved. While the IQE does not consider bias as a criteria for an acceptable quantitation limit value, this criteria could be added to the procedure.

2. Did the procedure meet the precision at L_Q established by the FACDQ?

The MQO established for precision at L_Q for the pilot study by the FACDQ was 20% relative standard deviation (RSD).

2.1. What is the data – What Works? What doesn't work? Confidence levels?

The pilot study results as related to the precision MQO are summarized graphically in the following attachments.
For Interlaboratory precision:

PSR.II.d.viii.1.a.(Interlaboratory Precision methods 300.0 and 335.4)
PSR.II.d.viii.1.a.(Interlaboratory Precision method 200.7) PSR.II.d.viii.1.a.(Interlaboratory Precision method 608) PSR.II.d.viii.1.a.(Interlaboratory Precision method 625)
For Single Laboratory precision:
PSR.II.d.viii.1.a.(Mean Laboratory Precision methods 300.0 and 335.4) PSR.II.d.viii.1.a.(Mean Laboratory Precision method 200.7) PSR.II.d.viii.1.a.(Mean Laboratory Precision method 608) PSR.II.d.viii.1.a.(Mean Laboratory Precision method 625)
For MMA PCB Study Precision:
PSR.II.d.viii.1.a.(Interlaboratory and Mean Laboratory Precision method 608)

Some description of the precision summary charts is necessary for the readers' understanding. Each interlaboratory precision chart includes four graphs representing the lab data;

- With Outliers (all data no statistical outliers removed) with precision estimated by modeling (Modeled)
- With Outliers with precision estimated by interpolation (Interpolated)
- Outliers Removed (statistical outliers removed from data) with precision estimated by modeling (Modeled)
- Outliers Removed with precision estimated by interpolation (Interpolated)

Each graph depicts the %RSD (identified by a small diamond) for each analyte of the method(s) specified by the title of the attachment. The X-axis lists analyte by each of the interlaboratory quantitation procedures evaluated. When a small diamond is either off scale or located on the zero %RSD axis it indicates that the procedure failed to yield a valid quantification limit estimate. Ideally if the procedure met the FACDQ pilot study MQO for precision all diamonds would fall at or below 20 %RSD. A diamond on the X-axis on the graph indicates that there was no variability at the given limit for that analyte.

The Mean Laboratory Precision graphs depicting single laboratory precision are formatted similarly to the interlaboratory precision graphs. Four graphs are also presented for each analytical method(s) evaluated. However instead of small diamonds the mean laboratory precision is identified by a small dash. For each analyte, error bars ranging from the minimum %RSD to the maximum %RSD, are also included. Ideally if the procedure met the FACDQ pilot study MQOs for precision all dashes and error bars would fall at or below 20 %RSD. If the procedures and analytical methods worked perfectly (all labs got the same perfect result) all dashes would fall on the 0% RSD line and no error bar would exist, indicating perfect precision with absolutely no variability from lab to lab.

RSDs observed in the Michigan Manufacturing Association (MMA) Polychlorinated Biphenyl (PCB) Study data by EPA method 608 are presented similarly to those described above. However, the X-Axis depicts data with outlier removal (OR), without outlier removal (WO), modeled (m) and interpolated (i). The average laboratory precision chart includes both PCB Aroclor 1016 and 1260, but the interlaboratory precision charts present each of the PCB Aroclors separately. The same rules apply when interpreting the results.

Data Analysis Findings

Outlier Removal for Interlaboratory Procedures – Across all methods and analytes outlier removal had minimal or no impact on how well the quantitation limit procedures performed in meeting the precision MQO. The exceptions to this rule include:

- Total cyanide by method 335.4, outlier removal produced both an IQE30 and IQE20 (interpolated) that achieved the precision MQO.
- Copper by method 200.7, outlier removal produced an IQE30 that achieved the precision MQO.
- Silver by method 200.7 (modeled), outlier removal produced an IQE30 that achieved the precision MQO.
- Endosulfan II and Endosulfan Sulfate by method 608, outlier removal improved precision for the IQE30.
- Benzo(A)pyrene by method 625, outlier removal allowed a valid LCMRL to be calculated, which achieved the pilot MQO for precision.

Outlier Removal for Single Laboratory Procedures – Across all methods and analytes outlier removal had minimal or no impact on how well the quantitation limit procedures performed in meeting the precision MQO. The exceptions to this rule include:

- Total cyanide by method 335.4, outlier removal resulted in a deterioration of average precision from near 20% to about 50% for SL-IQE20 and SL-IQE30. This appears to be due to one laboratory, which had several extreme values. While all of the extreme values in the limit calculation data were removed as outliers, not all of the extreme values in the limit confirmation data could be removed as outliers.
- Total Ortho-Phosphate by method 300.0 (interpolated), outlier removal allowed the precision MQO to be achieved for the ACIL-ML.
- Endrin II by method 608 (interpolated), outlier removal improved the precision of the LCMRL estimate.
- Endosulfan I by method 608 (interpolated), outlier removal improved the precision of the IQE20 estimate.
- Precision estimates for method 625 analytes improved sporadically with outlier removal. The most significant change was for Pentachlorophenol for the ACIL-ML, which dropped from about 40% RSD to below 20% RSD, with reduced variability between the individual laboratory RSDs. Because the ACIL ML doesn't target the lowest concentration with 20% RSD, the variability in lab RSDs could be due to variability in lab spike choices rather than variability of the limits.

Outlier Removal for MMA PCB Method 608 Data – Single laboratory precision variability for IQE20 and IQE30 was significantly improved from over 40% RSD to below 40% RSD for Aroclor 1016. An interlaboratory IQE20 limit could not be calculated after outlier removal for PCB Aroclor 1260.

Modeling versus Interpolation – In general, there was minimal observable effect whether the modeling or interpolation techniques were used to evaluate performance of the procedures. The only exceptions to this rule are as follows:

- Total Cyanide by method 335.4 (with outliers), interlaboratory IQE20 precision estimate is closer to the MQO using modeling.
- Total Ortho-Phosphate by method 300.0 interlaboratory IQE30 precision estimate is closer to the MQO using interpolation.
- Endosulfan Sulfate by method 608 (outliers removed), interlaboratory IQE30 precision estimate is closer to the MQO using interpolation.
- Benzo(A)pyrene by method 625 (with outliers), interlaboratory IQE30 precision estimate is closer to the MQO using modeling.
- Bis(2-ethylhexyl)phthalate by method 625 (with outliers), interlaboratory IQE30 precision estimate is closer to the MQO using modeling.

- Single laboratory precision estimates improved for methods 300.0 and 335.4 when modeling for the following parameters and procedures, Total Ortho-Phosphate (ACIL-ML), Total Cyanide (LCMRL and IQE20) and Nitrite (IQE20 and IEQ30).
- The single laboratory precision estimate improved for method 200.7 when modeling for Potassium when using the ACIL-ML procedure.
- The single laboratory precision estimate improved for method 200.7 when interpolating for Lead when using the IEQ30 procedure.
- The single laboratory precision estimate improved for method 608 when interpolating for Alpha-Chlordane when using the IEQ30 procedure.
- The single laboratory precision estimate improved for method 608 when modeling for Beta-BHC when using the IEQ30 procedure.

Generally, the difference between modeled and interpolated RSDs will be greatest when there is a large drop in RSD between consecutive spike levels. Linear interpolation predicts a much slower decrease in RSD compared to the modeled RSD. Based on the general RSD vs. concentration relationship throughout the concentration range, the modeled relationship is likely more reliable estimate.

General Observations

The inability to calculate a valid limit value (primarily the LCMRL with Methods 608 and 625) was discussed under the bias MQO and will not be repeated here.

Interlaboratory Precision – All procedures evaluated, the LCMRL, IQE20 and IQE30, consistently obtained limit estimates with precision of less than 40% RSD except for the following exceptions:

- Beryllium precision by method 200.7 fell between 40% and 50% RSD for IQE30.
- Lead precision by method 200.7 (interpolated) was approximately 50% RSD for IQE30.
- Endosulfan II precision by method 608 was approximately 60% RSD for IQE30.
- Endosulfan Sulfate precision by method 608 (interpolated) was approximately 50% RSD for IQE30.
- Pentachlorophenol precision by method 625 was ranged between 40% and 50% RSD for IQE30.

Precision below 20% RSD was achieved about half of the time, especially once statistical outliers were removed. Method 625 typically had precision at the quantitation estimate between 20% and 30% RSD. None of the procedures tested consistently achieved quantitation limit estimates with precision of less that 20% RSD.

For Methods 608 and 625, many analytes failed to yield an RSD below 20% at any concentration in the study. This tended to be due to either a laboratory having non-detects throughout much of the concentration range, and/or high biased results throughout much of the concentration range.

Single Laboratory Precision – The average laboratory precision MQO of 20% RSD was achieved or exceeded (less than 20% RSD) for all, methods and analytes when using the ACIL-ML and LCMRL procedures with the exceptions noted below.

- Average precision for Total Cyanide by method 335.4 was 30% - 40% RSD for the ACIL-ML and just barely over 20% for the LCMRL.
- Average precision for Potassium by method 200.7 (interpolated) was 30% RSD for the ACIL-ML.

- Average precision for Endosulfan I and Endosulfan II by method 608 was over 40% RSD for the ACIL-ML.
- Average precision for 2,4-Dinitrophenol and 3, 3'-Dichlorobenzidine by method 625 was about 40% RSD for the ACIL-ML.

The average laboratory precision MQO of 20% RSD was achieved for nearly all methods and analytes when using the IQE20, but not for the IQE30 which targets 30% RSD. The mean and median for all analytes was closest to 20% for the IQE20, because predicting the 20% RSD is the primary objective of this procedure. When a resulting limit value has an RSD of greater than 20% the limit will be lower than one that has 20% RSD. When a resulting limit value has an RSD of less than 20% the limit will be higher than one that has 20% RSD. The IQE generally showed greater variability in the %RSD between laboratories than did the ACIL-ML or LCMRL.

MMA PCB Study Method 608 - The LCMRL and IQE20 achieved the pilot study precision MQO after statistical outliers were removed for both single and interlaboratory procedures. The IQE30 achieved estimates with a precision between 20% and 30% RSD after outlier removal.

2.1.A. If it fails; why?

Interlaboratory Procedures

OGWDW LCMRL

Although the LCMRL procedure accounts for precision when generating a LCMRL value it does not target precision as one of the performance criteria it must achieve to calculate a valid limit. However, an LCMRL cannot be determined if there is large variability at a given spike level, even if there is little or no bias. Never-the-less the LCMRL achieved the pilot study precision MQO except for three occasions, (when a valid limit could be obtained). The three exceptions were all from EPA Method 200.7 and included Calcium (with outliers), Lead (interpolation with outliers) and Manganese. These results may be due to the inaccuracy of the interpolation or model used.

ASTM IQE

The ASTM procedure specifically targets a precision value (20% RSD for IQE20, 30% RSD for IQE30 and so on) in order to obtain a valid IQE. Because the IQE procedure targets the lowest concentration to achieve a specified RSD, on average, you would expect the estimated RSD at the resulting limit to exceed the target value half the time, and fall below the target value half the time. Therefore it is not surprising that the estimated RSD and the IQE20 sometimes exceeded 20%. The mean/median RSDs at the IQE20 were below 20% for all methods other than 625.

How close an IQE*N* will come to having *N*% RSD will depend upon how well both the data set and model used to calculate the IQE*N* represents actual performance of the population. As long as the requisite level of precision was achievable within the concentration range evaluated the IQE20 nearly always generated an RSD close to the precision MQO of 20%. If an IQE is unable to achieve the targeted %RSD this can be usually be corrected by collecting a dataset which better represents the entire population. However, in situations where the %RSD is highly variable in the quantitative range of the Method it is possible for a calculated IQE*N* to have more than *N*% RSD at concentrations greater than the IQE*N* (see Total Cyanide attachment below). Analyte/Method combinations were the IQE20 was unable to achieve the 20% RSD MQO included:
Method 300.0 for Chloride with outliers removed
Method 335.4 for Total Cyanide with outliers

Method 200.7 for Beryllium and Manganese (interpolation with outliers)
Method 608 for Endosulfan Sulfate (with outliers)
Method 625 for 2,6-Dinitrotoluene, Benzo(A)pyrene and Bis(2-ethylhexylphthalate)

PSR.II.d.viii.1.a.i.(IQE for Chloride by 300 with Low Level Spike)
PSR.II.d.viii.2.a.i.(IQE20 for Total Cyanide, entire population)
PSR.II.d.viii.2.a.i.(IQE20 for Beryllium, entire population)
PSR.II.d.viii.2.a.i.(IQE20 for Maganese, entire population)
PSR.II.d.viii.2.a.i.(IQE20 for Endosufan Sulfate, entire population)
PSR.II.d.viii.2.a.i.(IQE20 for 2,6-Dinitrotoluene, entire population)
PSR.II.d.viii.2.a.i.(IQE20 for Benzo(A)pyrene, entire population)
PSR.II.d.viii.2.a.i.(IQE20 for Bis(2-ethylhexylphthalate), entire population)

Single Laboratory Procedures

ACIL-ML

Although an ACIL-ML was always calculated it did not always achieve its precision objectives. This was a shortcoming of the pilot study itself, since it did not allow enough time for spike concentrations to be adjusted and reiterated. Method 200.7 precision failure rates are listed below for the ACIL procedure.

Method	>20% RSD
300.0	0%
200.7	3.4%
335.4	14%
608	20%
625	26%

The ACIL-ML procedure was written specifically to achieve the pilot study MQOs for individual laboratories and based upon verification sample performance it consistently achieved or exceeding (less than 20% RSD) the precision MQO of 20% average RSD. When the ACIL-ML did not achieve the target precision MQO was due to the presence of strong outliers for the laboratories which failed, as discussed previously.

The event that one spike level was attempted, and all replicates were detected but precision or accuracy MQOs were not met, the laboratories were instructed to try one higher level. If the MQOs were not met at this level either then they were instructed to proceed anyway. The intent of this direction was to prevent the labs in the pilot study from continuously having to try different spiking levels for method/analyte combinations that would not meet the precision/accuracy MQOs at any concentration. In addition, some laboratories did not consistently follow the direction to use a higher spiking level if the initial set of 7 replicates failed the precision/accuracy MQOs.

OGWDW LCMRL

The OGWDW LCMRL achieved or exceeded (less than 20% RSD) the precision MQO of 20% average RSD for all Analyte/Method combination, with the exception of Endrin by Method 608.

ASTM IQE

The IQE20 and IQE30 was unable to achieve their target %RSD on both an average and individual laboratory basis when the %RSD was highly variable in the quantitative range of the

Method for at least one laboratory. Under this set of conditions it is possible for a calculated IQE to have more than the targeted %RSD at concentrations greater than the IQE. Examples of this situation for several Method evaluated are include in the following attachments.

PSR.II.d.viii.2.a.i.(%RSD vs Conc T- Cyanide Lab 28)
PSR.II.d.viii.2.a.i.(%RSD vs Conc Gamma-Chlordane 29)
PSR.II.d.viii.2.a.i.(%RSD vs Conc 2,4-Dinitrophenol Lab 42)

3. Did the procedure meet the false positive rate for L_C as established by the FACDQ?

The MQO for false positive error rate at L_C for the pilot study established by the FACDQ was 1% or less.

3.1. What is the data? What Works? What doesn't work? Confidence levels?

The pilot study results as related to the false positive MQO are summarized in tabular form in the following attachments.
For Interlaboratory False Positives Error Rate;
PSR.II.d.viii.3.a.(False Positive Rates Interlab OR) PSR.II.d.viii.3.a.(False Positive Rates Interlab WO) PSR.II.d.viii.3.a.(High False Positive Rates Interlab)
For Single Laboratory False Positive Error Rate;
PSR.II.d.viii.3.a.(False Positive Rates Single Lab OR) PSR.II.d.viii.3.a.(False Positive Rates Single Lab WO) PSR.II.d.viii.3.a.(High False Positive Rates Single Lab)

Data Analysis Findings

Interlaboratory Performance – Out of 55 analyte-method combinations eleven (20%) did not achieve the false positive MQO for any procedures evaluated after statistical outliers were removed. Prior to the removal of statistical outliers twenty-six (47%) did not achieve the false positive MQO for any procedures evaluated.

Four of the 55 analyte-method combinations had false positive error rates greater than 10% for all procedures evaluated. Three of the four failed analyte-method parameters were common to datasets before and after outlier removal (Copper and Potassium by method 200.7 and Heptachlor by method 608). Di-N-Butyl Phthalate by method 625 failed prior to outlier removal and Bis(2-Ethylhexyl) Phthalate by method 625 failed after outlier removal. The Drinking Water interlaboratory Hubaux-Vos procedure had a slightly higher failure rate with Di-N-Octyl Phthalate and Chrysene also failing for method 625.

Single Laboratory Performance – Out of 374 analyte-method-laboratory combinations an average of 111 (30%) did not achieve the false positive MQO for the procedures evaluated (when including limits with and without outlier removal as separate evaluations). The percentage of analytes/laboratories that yielded false positive rates of 1% or below for each of the single laboratory limits were:

	With Outliers	Outliers Removed
ACIL MDL	65%	72%
ASTM SL-Yc	67%	78%

ASTM SL-Lc	71%	78%
DW-HV Yc	61%	71%

Only twelve of the 374 analyte-method-laboratory combinations yielded false positive error rates greater than 10% for all evaluated single-laboratory detection limits. Seventy-one (19%) analyte-method-laboratory combinations were unable to achieve the false positive MQO with one or more of the limits evaluated. The majority of the 71 high false positive rates were from five laboratories performing either analytical method 200.7 or 608. Statistical outlier removal had a minimal effect on extreme false positive error rates; however the effect was larger for a subset of the analyte-method-laboratory combinations.

3.1.A. If it fails; why?

Interlaboratory Procedures

Generally speaking the false positive error rate MQO of 1% was exceeded when a laboratory contributed a large percentage of the total blanks and an even larger percentage of the high blanks (those exceeding the Lc) While an equal number of results per laboratory was used to calculate the limits, the number of blanks per laboratory used to assess the limits differed. If the number of blanks, and the blank distribution, varied widely between labs, the 1% MQO was generally not met. Examples of this include; Lab 8 for Copper by method 200.7, Lab 4 for Potassium by Method 200.7 and Lab 32 for Heptachlor by Method 608.

There were often differences between the ASTM-Lc and the Yc calculated by both the ASTM and OGWDW IL-HV procedures. The ASTM Lc includes a recovery correction, unlike the two other interlab detection limits. For methods such as 625, with low recoveries, this means that the ASTM Lc will tend to be higher. However, for Method 300.0, there were often very high recoveries at one or more of the calculated spike levels. Therefore, the recovery correction would mean that the ASTM Lc would be lower than the other limits. Examples of this include; Fluoride and nitrite by Method 300.0 (high recoveries) and Di-N-Butyl Phthalate by Method 625 (low recoveries).

Single-Lab Procedures

Generally, the false positive rates were lowest for the ACIL MDL for uncensored methods. For these methods, the ACIL MDL is calculated from blank results, and does not extrapolate down based on spiked data. The ASTM and OGWDW detection limits were calculated using spike levels originally chosen by the labs to target quantitation limits. The effect of this choice is assessed in Section II d iv. The regression equations used by the ASTM and OGWDW procedures do not always predict the actual performance of the method blanks, because the fitted recovery regression intercept does not accurately estimate the blank bias To further complicate the issue the historic blanks did not always have the same false positive error rate as the blind and QC blanks analyzed during the pilot study. For, example, there were instances when the calibration blanks were higher than the routine laboratory blanks for Method 200.7.

For censored methods, the lowest false positive rates occurred for the ACIL MDL for Methods 608 and 625, but not for Method 300.0. For censored methods, the ACIL procedure generally assumes no bias in blanks. The procedure states that if frequent detects are observed in blanks, the MDL should be calculated following both the uncensored and censored limit procedures, and the higher MDL should be used. However, for many laboratories and analytes, the majority of blanks did not yield detects but the mean of the blanks was still well above 0. In these cases, it is not known whether laboratories would apply the uncensored procedure. The censored ACIL

procedure does not include any adjustment for blank bias which can result in a high false positive error rate (Lab 13 for Sulfate by Method 300.0). The ASTM Hubaux-Vos detection limits do not assume the mean blank concentration is equal to zero, but instead estimate the mean blank based on the spiked data. Generally, these estimates tended to be yield more reliable false positive rates for Method 625 compared to Methods 608 and 300.0.

Among the ASTM detection limits, the false positive rates tended to be slightly higher for the SL-Lc compared to the SL-Yc. This difference depended mainly on the recoveries observed in the spiked sample results used to calculate the limits. This is discussed further in Section 4, below.

To summarize, the key cause of differences between actual blank false positive error rate and regression based predictions include the following:

- Differences between laboratory blank contamination and calibration blank contamination or sample contamination or study blank contamination.
- Extreme outlier effect on either the calculated limits or the blank distribution.
- Very precise spike data (not representative of long term laboratory performance) leading to low regression predictions, resulting in an underestimation of the blank variability.
- Intermittent blank contamination, resulting in parametric procedures not able to adequately predict performance.

For these reasons is it important that a non-parametric intermittent blank contamination procedure be incorporated in any detection procedure and that laboratory blank performance be monitored on an on-going basis in order to adjust the detection limit estimate over time as needed. While the majority of the high false positive error rates were due to the performance of the estimation method in light of the highly variable blank data, some of the failures were due to the incorrect application of the procedures by some of the laboratories in the study. Another shortcoming was the limited time period over which the pilot study was completed, because this in some cases did not allow an adequate time to incorporate actual laboratory performance in the estimates.

4. Did the procedure meet the false negative rate at L_C for the true value at L_D or L_Q as established by the FACDQ?

The MQO for false negative error rate at L_C for results at the L_Q for the pilot study established by the FACDQ was 1% or less.

4.1. What is the data? What Works? What doesn't work? Confidence levels?

The pilot study results as related to the false negative MQO are summarized in tabular form in the following attachments.
For Interlaboratory False Negatives Error Rate;
PSR.II.d.viii.4.a.(False Negative Rates Interlab 200.7) PSR.II.d.viii.4.a.(False Negative Rates Interlab 300.0) PSR.II.d.viii.4.a.(False Negative Rates Interlab 335.4) PSR.II.d.viii.4.a.(False Negative Rates Interlab 608) PSR.II.d.viii.4.a.(False Negative Rates Interlab 625)
For Single Laboratory False Negative Error Rate;
PSR.II.d.viii.4.a.(High False Negative Rates Single Lab 200.7) PSR.II.d.viii.4.a.(High False Negative Rates Single Lab 300.0) PSR.II.d.viii.4.a.(High False Negative Rates Single Lab 335.4) PSR.II.d.viii.4.a.(High False Negative Rates Single Lab 608) PSR.II.d.viii.4.a.(High False Negative Rates Single Lab 625)

Data Analysis Findings

Interlaboratory Performance – Out of 55 analyte-method combinations the ASTM procedures (the detection decision made at the Yc or Lc, with the false negative rate evaluated at the IDE, IQE20 or IQE30) was unable to meet the false negative error rate of one percent the majority of the time. The only exception to this was for Methods 608 and 625, where the Yc and Lc evaluated against the IQE20 achieved the false negative error rate objective the majority of the time. However the failure rate for the ASTM procedure was 33% of the labs/analytes for method 608 and 17% of the lab/analytes for method 625 based on Lc. It is not surprising that the IDE did not achieve the 1% FN MQO, because this procedure was written to achieve a 5% FN error rate. It is also not surprising that the IQE30 did not achieve the MQO, because the IQE30 tended to fall below the IDE. The HV-Yc evaluated against the LCMRL achieved the target MQO of one percent false negatives the majority of the time; however the LCMRL procedure was unable to produce a valid limit the majority of the time for Methods 608 and 625, due primarily to the high variability of these methods. Statistical outlier removal slightly improved the ability of a procedure to achieve the established false negative MQO.

Single Laboratory Performance – False negative error rates varied by procedure (ACIL-MDL, ASTM SL-Yc, ASTM SL-Lc, and HV-Yc), evaluation limit (ACIL-ML, ASTM SL-IDE, ASTM SL-IQE20, ASTM SL-IQE30 and LCMRL), analytical method (300.0, 335.4, 200.7, 608 and 625) and analyte. To a lesser extent, the method used to estimate the false negative rate (i.e., linear interpolation or modeling) and the inclusion/exclusion of outlying results also impacted the error rate. The best performance was the ACIL-MDL evaluated against the ACIL-ML for Method 300.0, where a false negative error rate of less than one percent was achieved nearly 100% of the time. The worst performance was the ASTM-Lc evaluated against the IQE30 for Method 625, where the false negative MQO was missed nearly 100% of the time. Overall the single laboratory false negative error rates followed the general rules of thumb.

Highest %FN ASTM SL-Yc > ASTM SL-Lc > HV-Yc > ACIL-MDL Lowest %FN

Highest %FN SL-IQE30 > SL- IDE > SL-IQE20 > LCMRL > ACIL-ML Lowest %FN

Highest %FN 625 > 608 > 200.7 > 300.0 Lowest %FN

Highest %FN Interpolation > Modeling Lowest %FN

Highest %FN With Outliers > Outliers Removed Lowest %FN

Exceptions to this rule include method 335.4 (Total Cyanide) which only achieved the target MQO about half the time no matter which procedure was employed. The LCMRL also nearly always achieved the target MQO for Methods 608 and 625; however a valid LCMRL could not be calculated for approximately 50% of the labs and analytes for these methods. It should also be noted that sporadic high or extremely high false negative error rates do occur. This will be discussed when we look at the cause of high false negative error rates.

Interpolated false negative rates tended to differ greatly from the modeled false negative rates due to the large drop in the observed rates between consecutive spike levels. Linear interpolation predicts a much slower decrease in false negative rate compared to the modeled rate. Based on

the general false negative rate vs. concentration relationship throughout the concentration range, the modeled relationship likely yields the more reliable estimate.
While high outlying results can have a strong effect on the calculated limits, their effect on the confirmation assessments will be mitigated. This is because each result is categorized as a detect or a non-detect. Therefore, a result that slightly exceeds a specified detection limit will have the same effect on the model or interpolation as a result that greatly exceeds a specified detection limit. Low outlying results, especially at higher spike levels, will have a greater effect on the confirmation assessments. This is because a single low outlier that falls below the evaluated detection limit will result in a false negative rate of at least 10% at that spike level, which will inflate the modeled or interpolated false negative rate.

Focusing on the more extreme false negative error rates, out of 1,870 analyte-method-laboratory-evaluation limit combinations the false negative error rate exceeded 10% 886 times (47% of the time). The percentage of time the false negative error rate exceeded 10% for all procedures (prior to outlier removal) is listed below by analytical method.

- Method 300.0 27%
- Method 608 39%
- Method 200.7 55%
- Method 625 58%
- Method 335.4 62%

In many evaluations, the interpolated rate exceeded 10% while the modeled rate fell below 10%. Based on the above discussion, it is likely that the modeled rate is the more reliable estimate. Further discussion of the specific procedures is presented below:

ACIL

For censored methods, the ACIL MDL is calculated based on the variability of the results of samples spiked at the ML. Therefore, the false negative rate is heavily influenced by the recovery and RSD observed in those spiked sample results. For the laboratory-spiked data used to set the ACIL ML and calculate the ACIL MDL tended to be very precise and accurate. Therefore, the calculated MDL was well below the chosen ML spike level, and the false negative rates tended to be very low. For Methods 608 and 625, the recoveries of samples spiked at the ML tended to be much lower, and the RSD tended to be much larger**.** As a result, the resulting MDL was much closer than the ML, and the false negative rates were larger. In many of these cases, the mean recovery and RSD failed the Pilot Study and procedure MQOs; in practice the ML and MDL would be re-evaluated at a higher spike level. The effect of this re-evaluation would depend on how the recovery and variability change based on the increased spike level. Also, many cases of high false negative rates in the ACIL procedure were due to laboratories failing to follow the direction that the quantitation limit must be elevated to at least 2X the detection limit.

ASTM

The false negative rate tended to exceed 10% most frequently for the SL-IQE 30%. For the majority of analytes and laboratories, the observed RSD at spike levels around the calculated detection limits was closed to 30%. As a result, the calculated IQE 30% was often approximately equal to, and sometimes below, the calculated detection limits. As a result, the false negatives tended to be high at the calculated SL-IQE 30%.

For each of the single-laboratory quantitation limits, the false negative rates tended to be higher when detection was based on the SL-Lc compared to the SL-Yc. This was because the calculated SL-Lc tended to be higher than the calculated SL-Yc. The ASTM SL-Lc calculation includes a recovery correction. For Method 625, which had low recoveries for most analytes, the recovery correction increased the Lc to a level approximately equal to the IQE30.

Generally, the false negative rates were slightly higher for the SL-IQE20 compared to the SL-IDE. The calculated SL-IDE tended to be slightly higher than the calculated SL-IQE20. This difference depends not only on how recovery and variability change with increasing concentration, but also on the amount of data used to calculate the limits. If the SL-IDE and SL-IQE20 are determined using similar standard deviation vs. concentration models, the difference between the SL-IDE and SL-IQE will be based almost entirely on how many results were used in the limit calculations. This is because the number of results will directly affect the SL-IDE, which is based on two tolerance multipliers determined based on the number of results, but not the IQE20. For the single-lab limit calculations performed using single-lab data, the tolerance limits will result in an SL-IDE greater than the SL-IQE20, when both are determined using the same models. Therefore, because the SL-IDE would be expected to yielded false negative rates closer to 5% than 1%, the SL-IQE20 also tended to yield false negative rates greater than 1%. In the interlaboratory case, more results were used to calculate each IDE, and therefore the tolerance limit multipliers decreased. As a result, the interlaboratory IQE20 was less likely to be below the interlaboratory IDE, and the false negative rate at the IQE20 was less likely to exceed the false negative rate at the IDE.

OGWDW

Compared to the other procedures, the LCMRL procedure tended to yield a quantitation limit much greater difference between the determined limit and its associated detection limit. This is due to the much stricter target MQO for the LCMRL compared to the other limits. This stricter MQO resulted in the inability of the procedure to calculate a valid LCMRL for a large number of analytes/labs. An assessment of whether the LCMRL calculation failed based on laboratory spike level choice or the method/lab performance is found in Section II d iv.

4.1.A. If it fails; why?

Interlaboratory Procedures

OGWDW IL-HV Yc

The false negative rate at the IL-HV Yc was nearly always less than 1% for LCMRL values that were calculated. Out of the four instances when the false negative error rate was greater than 1% on only one occasion was the error rate much greater than 2%. Manganese by Method 200.7 had a FN error rate (80% modeled rate and 75% interpolated rate, both without outlier removal) and the IL-HV Yc calculated for Manganese was actually higher than the IL-LCMRL, due to the presence of a couple of high recoveries at the lowest spike level (resulting in a high recovery intercept). When the detection limit exceeds the quantitation limit, the false negative rate will normally be high. For example, assuming an accurate result (very precise with minimal bias) at the QL, were the DL = QL and the FP error rate at the DL is 1%, we would expect a FN error rate of 50%.

ASTM (Yc, Lc, IDE, IQE)

The false negative rate at Yc and Lc were nearly always greater than 1% for the IQE30 and often greater than 1% for the IQE20. In general the closer the quantitation limit comes to the detection limit the higher the false negative error rate, though the extent of this differs between methods. The reasons that the pilot MQO for false negatives was sometime not achieved by the ASTM procedure were two fold.

1. The ASTM procedure is designed to return a false negative rate at the Lc for values at the IDE of 5%. Therefore when the IDE = IQE the FN error rate should be 5% in theory. Approximately 60% of the time the IDE was greater than the IQE30 and approximately 40% of the time the IDE was greater than the IQE20, neither of which was unexpected. This means that the floor of 5% FN would often be the driver setting the IQE at the IDE. If the FN rate at the IDE were set at 1% this would bring the ASTM procedure much closer to the pilot FN rate MQO. Another option would be to select a higher precision MQO, such as 10% since an IQE10 should never fall below the IDE.

2. Less frequently, high recoveries at the lowest spike levels can result in a high recovery intercept, resulting in an IQE that falls well below the Yc or Lc. For example, Chloride by method 300.0 and Bis-2-ethylhexylphthalate (with outliers), both exhibited this phenomenon.

Single-Laboratory Procedures

ACIL-MDL

Overall the false negative rates at the ACIL-MDL, for values at the ACIL-ML were well below 1%. This is because the further the quantitation limit is from the detection limit the lower the false negative rate and the ACIL procedure incorporates a requirement that the QL be at least 2x the DL for uncensored methods. The occurrences where the FN rate was greater than 1% were most often a direct result of the laboratory not following this requirement in the ACIL procedure. For Methods 335.4 and 625, the FN rate exceeded 1% for some labs and analytes; this sometimes occurred because the ongoing spikes had > 20% RSD, which in practice would have lead to an increase in the QL when following the ACIL procedure. Unfortunately, the pilot study time period did not allow an evaluation of this portion of the procedure.

OGWDW HV Yc

The false negative rate at the Yc was nearly always less than 1% for LCMRL values that were calculated. In every instance when the false negative error rate was greater than 1% the HV Yc was greater than the LCMRL. For example; Be, Cd, Cu, Pb, Mn and Zn by lab 8 (Method 200.7), Chloride by lab 17 (Method 300.0) and Endosulfan Sulfate by lab 32 (Method 608). As with the interlaboratory estimates when the HV Yc is greater than the LCMRL this is normally a result of high recoveries at the lowest spike levels resulting in a high recovery intercept.

ASTM (Yc, Lc, IDE, IQE)

The false negative rate at Yc and Lc were nearly always greater than 1% for the IQE30, with the mean FN error rates approaching 50%. The false negative rate at Yc and Lc were often greater than 1% for the IQE20. In general the closer the quantitation limit comes to the detection limit the higher the false negative error rate. The reasons for failure of the pilot MQO for false negatives when applying the ASTM procedure to single-laboratories are directly related to the procedure objectives as discussed under the interlaboratory section. When the detection limit exceeds the quantitation limit, the false negative rate will normally be high. For example,

assuming an accurate result (very precise with minimal bias) at the QL, were the DL = QL and the FP error rate at the DL is 1%, we would expect a FN error rate of 50%.

ix. Implications of Non-Normally Distributed Data to Measurement Quality Objectives (MQOs). (TF/RR/KO)

1. Causes of Non-Normality Near the Detection Estimate

Ideally, analytical measurements are impacted only by random errors with repetitive measurements resulting in data sets described by normal (Gaussian) population distributions. Systematic errors overlaid onto random errors can result in skewed or kurtotic distributions (positive or negative), especially at concentrations where measurement sensitivity is challenged and systematic errors represent a substantial fraction of the analytical signal. Experience has shown that analytical measurements performed on samples having analyte concentrations near or below the critical level often result in data sets that appear to be non-normal (non-Gaussian). Such distributions sometimes appear log-normal, although it has been argued that characterizing the distributions would require exceeding large numbers of measurements (see for example, *The Frequency Distribution of Analytical Error,* Analyst (1980) 105, 1188 – 1195). Nevertheless, data censoring thresholds and analytical artifacts (e.g., spurious contamination and analyte losses) play a role in how data appear to be distributed. As described below, there are potential implications of non-normal data distributions when attempting to evaluate measurement quality objectives such as false positive and false negative rates for detection procedures developed around the premise of normality.

The pilot study employed two general types of analytical methods: techniques that produce a quantitative response for every measurement regardless of analyte concentration (uncensored methods) and techniques that produce no quantitative response below a certain signal threshold (censored methods), as described in section II (d) (iv), above. As expected, data censoring can have profound impacts on the resulting distribution. For example, only about 7 % of the 4,4-DDT method blanks submitted by one laboratory participating in the pilot study were reported quantitatively. The other 93 % of the method blank results were censored. Similar trends were apparent for other analytes across most laboratories and methods where censoring techniques were employed. In some cases, censoring was virtually complete for method blanks, as expected. Where censoring is employed in the signal or concentration domain, positively skewed data distributions often result at low concentration ranges around the censoring threshold, primarily because data on one side of any underlying distribution are censored. An evaluation of representative pilot study data submitted for censored methods demonstrated that the condition of normality was often rejected for method blank data sets. If censoring could be 'turned off', methodological noise distributions thus resulting might reveal noise distributions hidden by censoring that were not as skewed as when viewed through a censoring filter. In addition to hiding the nature of the methodological noise distribution (key to establishing detection limits for some procedures), the central tendency (of the methodological noise distribution) cannot usually be ascertained within full or partially censored data sets.

Non-normal data distributions can also result for non-censored methods and are often observed for trace analytical methods or analyses where laboratory or reagent contamination is likely. In those cases, the methodological noise distribution will be positively skewed and the central tendency of the distribution will be greater than zero. When the laboratory method blank data associated with the pilot study were evaluated, those analytes typically associated with laboratory

contamination events (such as aluminum, iron, copper and zinc) often had non-normal distributions as depicted below.

See Cumulative frequency distributions for Al (left) and Zn (right) from pilot study method blank results for the laboratory submitting the largest amount of data (> 1000 points). Normality would be indicated by points falling on the line in appendices.

Of the eight laboratories that submitted method blank data sets for those elements by Method 200.7, 63% of the aluminum data sets 75% of the copper data sets, 75% of the iron data sets and 75% of the zinc data sets were non-normally distributed. For arsenic, where positively skewed data resulting from laboratory contamination are less likely, only 25% of the laboratory method blank data sets were non-normally distributed. However such generalizations should be avoided; 100% of the method blank data sets submitted for the pilot study were non-normally distributed for cyanide despite the fact that cyanide is not a typical laboratory contaminant (50% of the cyanide data sets exhibited positive skewness or positive kurtosis).

2. Implications of Non-Normal Data Distributions on Measurement Quality Objectives for Detection and Quantitation Procedures

Virtually all the procedures tested in the pilot study make some initial assumption that the data sets used to calculate detection and quantitation metrics are normally distributed. Those assumptions are described explicitly (such as in section 6.2.4 of the ASTM IDE/IQE procedures) and implicitly through the use of statistical parameters used to derive detection estimates (such as the 'k' statistic and Grubbs employed in the ACIL procedure, or in the case of Hubaux-Vos/LCMRL, by assuming normality in the residuals). Non-normal or skewed data distributions can have implications to most measurement quality objectives for detection procedures that assume only random errors will affect measurements. Systematic errors leading to non-normal distributions are likely to have the greatest effects on measurement quality objectives in the region around the detection estimate where the magnitude of systematic errors is more comparable to that of random errors. Larger systematic errors that could potentially affect analytical precision and recovery at and above quantitation limits are more likely to be recognized and corrected, where feasible. Furthermore, some of the procedures tested in the pilot study incorporate adjustments to quantitation limits in order to ensure recovery and precision objectives are satisfied.

For reasons discussed above, false positive rates and false negative rates are the two measurement quality objectives most likely to be impacted by non-normal data distributions when normality is a procedural assumption. The nature and extent of impacts will be dependent on the type of distribution and the magnitude and direction in which data used to establish detection estimates may be skewed (kurtosis). Computer simulations using several normal and positively skewed data distributions revealed that in all cases tested, moderate positive kurtosis increased false positive rates (measured as the number of observations in the data set exceeding the calculated detection estimate) for non-regression techniques such as the ACIL procedure. A representative simulation is shown below.

See Effects on the detection estimate and the false positive rate of positively skewing a normal data distribution

In all cases simulated where the distribution was positively skewed, the detection estimate (L_C) moved in a positive direction (toward higher concentration) but not sufficiently to exclude the

expected number of false positives (≤ 1%). However, the false negative rate measured at an independently derived quantitation limit (L_Q) also increases for positively skewed distributions because L_C shifts positively.
Many of the procedures tested in the pilot study yielded false positive and negative rates higher than predicted by the procedures. Given the relatively small method blank data sets, if procedures had performed as designed few or no false positives should have been observed when that data were compared against calculated detection estimates. There are a number of reasons why the procedures may not have performed as expected, however the effect of non-normal method blank data sets cannot be ignored as a contributing factor. Incorporating non-parametric tests to discard outliers may improve the performance of the procedures.

Another interesting phenomenon was observed with the pilot study data that can affect false positive and false negative rates. For procedures that rely on the extrapolation of data collected at higher spike concentrations to establish detection estimates, there is a real possibility that the spiked data sets may have distribution characteristics different from the distribution that might exist at the detection estimate. During the pilot study, fortified samples (single-blind spikes) across a concentration gradient were submitted to laboratories and the results were used to determine detection and quantitation limits for a number of procedures that rely on extrapolation techniques (Hubaux-Vos, LCMRL, ASTM IDE/IQE). When those pilot study data for representative 200.7 analytes were evaluated, it was observed that distributions at the higher spike concentrations data were more often normally distributed than data for the lowest spike concentrations. A summary of that evaluation is depicted below.

See Pilot study data sets passing (P) and failing (F) a normality test for representative analytes in Method 200.7 in appendices.

This table shows how the assumptions made during the determination of the detection estimate may differ from the actual characteristics of the data used in the calculation, especially for single laboratory procedures relying on extrapolation techniques. Pooled data used for interlaboratory detection and quantitation estimates show similar trends. Based on data collected for the pilot study, it appears that violation of the normality assumption inherent in the tested procedures may account for some of the observed failures to achieve the stated measurement quality objectives.

x. Evaluate pilot study data and other procedures not pilot tested (CG & LabQC). (JP/KO/SW)

Consensus Group Lc and QL Procedure
The Consensus Group Lc and QL procedure is very similar to the ACIL MDL and ML procedure, except it has more detailed instructions and guidance. The key differences between the procedures are identified in the following attachment.

PSR.II.d.x.(Key Differences Between ACIL and CG Procedures)

In general the Consensus Group (CG) procedure provides slightly tighter (+) controls on the limit values generated, but in some cases may be less stringent (-) including the following:

+ CG procedure has more clearly defined criteria and more stringent controls on Quantitation Limit Check (QLC) sample

- \+ CG procedure has an additional QL criteria - standard deviation must be less than 20% of the QLC spike concentration
- \+ CG procedure specifies ongoing control charting of QLCs
- \+ CG procedure has more frequent verification of estimates
- \+ CG procedure has a provision to decrease estimate if the false negative error rate is below 1%
- \+ ACIL procedure has a provision to report a qualified result if precision and bias criteria can not be achieved
- – CG procedure allows the QLC to be 3 x Lq for up to 10% of analytes in a multi-analyte method
- – CG procedure recovery criteria can be relaxed to 10-190% for up to 20% of analytes in a multi-analyte method

Some of the key differences noted above and listed in the comparison table can be evaluated with existing pilot study data, but most can not. When the difference can be tested we will provide those results. For other differences we will provide a general discussion.

Tested Differences

Item one in the difference table "percentage of numeric results for uncensored procedure" can be evaluated from pilot study data by recalculating the MDL or Lc limits using blank datasets, which have between 85% and 100% numeric results. This was done and is presented in '*PSR.II.d.x.(Percentage of Numeric Results Difference Between ACIL and CG Procedures)*' . For censored methods, which frequently produce numerical results for the blank the ACIL procedure allows the MDL to be calculated using both blanks and spiked blanks, and defaults to the greater of the two. The pilot study did not evaluate this portion of the ACIL procedure so we will do so here. In general the MDL or ML derived using censored data did not vary more than 2-3 fold from the limits derived from uncensored data. The only exceptions to this was for Chloride by method 300.0 lab #16 and Di-N-Ocytyl Phthalate by method 625 lab #45. The uncensored limit value for Chloride lab #16, may be unrealistically low because exactly the same numeric value was often reported for blank results, which may be due to excessive rounding. The uncensored limit values derived from Di-N-Octyl Phthalate blank results for lab #45 were low due to very consistent contamination in the blanks. Seventeen percent of the limits calculated for Method 608 would have been calculated using the uncensored technique and would have yielded a higher limit value if the CG procedure was used. If the ACIL procedure was performed as written at least 17% of the limits calculated for Method 608 could have been calculated using the uncensored technique and would have yielded a higher limit value.

Difference item number ten "acceptance requirements for the QL", can be tested using pilot study data by evaluating the number of ACIL derived ML values, which passed the precision and bias criteria but fail to meet CG criteria of s/Lq x 100 < 20%. This was done and in no cases did a limit value pass the ACIL precision and bias criteria yet fail the CG bias criteria based on spike concentration. On several occasions the spike bias approached 20% but never exceeded 20%. High bias relative to spike concentration would be expected with a combination of high bias at the QLC and recoveries greater than 100%. Apparently this occurred very infrequently in the pilot study. A more effective means of reducing this type of error might be to put stricter controls on the upper percent recovery allowed. For example the recovery criteria could be adjusted from 50-150% to 50-120% or 50-130%.

Untested Differences

In general most differences between the ACIL ML/MDL and CG Lc/QL procedures would result in lower detection and quantitation estimates and estimates which are more representative of day to day laboratory performance. This is because of the stricter requirements/controls of the CG procedure and more frequent verification of estimates. This would come at a cost, because QLC standards would need to be run more frequently and additional time and effort would be required in producing the estimates. The only exception to this might be for certain poor performing analytes in multi-analyte methods such as method 608, 624 and 625. In this case the CG procedure relaxes the recovery criteria substantially to 10-190%.

Wide recovery acceptance criteria can have a direct impact on the false negative error rate. The false negative error rate increases as the QL (ML) approaches the DL (Lc or MDL). In both the ACIL and CG procedures the QL must be greater than or equal to twice the DL. With acceptance recovery criteria as low as 50% a QL (QLC) value of 20 could produce a reported result as low as 10. In this case the DL could be equal to the measured QL (QLC) and by definition the false negative error rate would be 50%, which is significantly greater than the 1% objective. Wider recovery acceptance criteria as proposed by the CG procedure (as low as 10% for some analytes) would increase the false negative error rate to be well over 50%. Based on these principles it would be much better (lower false negative error rates) if the QL was required to be at least three times the DL. This shows that the 3.18 multiplier used between the MDL and ML makes reasonable sense. While not advocating a fixed multiplier between the DL and QL, it is recommended that a minimum separation between the DL and QL of greater than 2x be required or the minimum percent recovery criteria be increased to 75%.

Consensus Group Lc and Ld procedure

The Lc derived from the CG Lc/Ld procedure is identical to the Lc derived from the CG Lc/QL procedure. The Ld derived from the CG Lc/Ld procedure is similar to the QL derived form the CG Lc/QL procedure. The key difference is that the QL must meet the precision and bias criteria established for the pilot study. Since these precision and bias criteria as so loose the Ld and QL very similar. Essentially the QL in the CG procedure is established with Ld as the lowest possible value for QL (Ld is the floor of the QL). So the QL may be greater than the Ld, but may never be lower than the Ld. Non-parametric on-going verification steps are built into all three procedures (ACIL MDL/ML, CG Lc/QL and CG Lc/Ld), which ensures that over time that the QL will never fall below the Ld and that the Ld will never have more than a one percent false negative error rate. Unfortunately, the ongoing verification procedures were never evaluated during the pilot study due to time constraints. Because the CG procedure requires more frequent verification than the ACIL procedure we would expect the CG would procedure produce estimates more representative of actual laboratory performance.

e. Conclusions and Findings. (RB/JP)

i. General Observations

The procedures that were tested in the Pilot Study were successful in that most of the MQOs were met most of the time. However, clear differences between the procedures emerged, and weaknesses of the procedures could be identified. At least some of these weaknesses are amenable to correction with relatively minor adjustments to the procedures.

There were large differenced between laboratories both in the ability to meet the MQOs and in the absolute levels of detection limits determined. This has serious consequences for the general applicability of inter-laboratory estimates such as the IDE/IQE.

The MQO that failed most often was the false negative rate. In part this is due to the fact that this MQO has two components, the detection limit and the quantitation limit. If the detection limit is too high relative to the quantitation limit, then both false positive and false negative rates will be elevated.

When false positive rates were high, the failing was often due to intermittent blanks (frequent blank result observations for analytical methods that were considered censored, i.e., not all of the blank had numerical results). The Intermittent Blank Contamination Guidance developed by TWG seems to address this adequately.
The accuracy MQO of 50-150% was generally met. When it was not, some of the cases could be identified as “poorly performing analytes”, those method/analyte combinations for which recovery might not achieve 50% for the entire analytical range. There were other cases where poor recovery would not be expected, yet the 50-150% criterion was missed on either the low or the high side. The cause of this phenomenon has not been thoroughly explored, but it may well be due to calibration error introduced by the use of unweighted linear regression for calibration curves. This type of calibration curve can often result in large relative error at the low end of the calibration.

In some cases laboratories did not closely follow the directions in the ACIL procedure and/or the Statement of Work. Laboratories that failed to follow the procedure often failed to meet the MQOs by a wide margin. For example, laboratories that did not follow the procedure’s instructions for choosing a spike level often determined MLs which failed the RSD and false negative rate MQOs. Additionally, laboratories were not instructed to follow certain parts of the ACIL procedure due to the necessity of compressing the data collection into a shorter timeframe than would be normal, which resulted in biased-low MLs in certain cases. However, it also indicates the need to make the language in the final procedure very clear, and also to provide support documents.

ii.Outlier Removal

Much of the data analysis was performed both with and without outlier removal. In general, the effects of outlier removal were slight and do not change the overall conclusions. This was likely due to outlier removal having two potential outcomes: reducing the bias and/or variability at a determined detection or quantitation limit, and reducing the calculated detection and/or quantitation limits themselves.

iii. Strengths and Weaknesses of Each Procedure

All procedures (when performed as written) failed to produce valid estimates when the analytical method was unable to achieve the pilot study MQOs at any concentration. This is strength, because it warns the user that data generated by the method to those parameters is not reliable.

All of the procedures failed to work effectively in meeting the false positive MQO when there were intermittent blanks. A task force has developed a set of recommendations for techniques of dealing with the intermittent blank case, and these concepts should be incorporated into the final procedure chosen.

Most of the procedures had difficulty with the false negative MQO in cases where the quantitation limit was too close to the detection limit. The ACIL procedure includes a requirement that the quantitation limit be at least two times the detection limit, but this specification was not followed reliably by the laboratories. The requirement needs to be strengthened in the ACIL procedure and incorporated into other procedures.

ACIL Procedure

Because this procedure is designed as a demonstration of performance rather than a determination of a lowest possible quantitation limit, this procedure tended to meet the study MQOs more frequently than other procedures. When the pilot study MQOs could be achieved at one of the concentration tested, most failures to meet the MQOs using the ACIL procedure were due to laboratories not following all of the requirements, in particular the requirement that the quantitation limit be at least two times the detection limit and the requirement respike at a higher concentrations if the MQOs for precision and accuracy were not met at the original concentration chosen. The ACIL procedure needs to be clarified and strengthened to avoid these errors, and intermittent blank guidance needs to be incorporated. Concepts from the Consensus group procedure and LABQC procedures also need to be considered for incorporation into the ACIL procedure.

LCMRL / Hubaux Vos

The Hubaux Vos procedure did not work too well, and the Office of Water has suggested that the LCMRL be considered as a quantitation limit procedure only and paired with something along the lines of the ACIL procedure for detection. The LCMRL procedure did not identify quantitation limits for a substantial number of method 625 analytes and some method 608 analytes. This is because even though this procedure determines a “lowest possible” quantitation limit, it is based on an MQO (prediction interval for 50-150% recovery of individual sample results) that is quite stringent compared to the study MQOs. If the LCMRL MQOs were relaxed, then limits would be obtained for more analytes. Alternatively, the analytes that do not meet the LCMRL criteria at any point in the calibration range could be consider as non- or semi-quantitative. This would unfortunately require difficult decisions regarding the uses of this data. If the LCMRL was paired with the ACIL procedure, the requirement that the quantitation limit be at least two times the detection limit would be vital in order to avoid high false negative rates for some analytes.
An unfortunate weakness of the Pilot study was that while the LCMRL procedure was evaluated, the MRL was not. The Office of Water intended that the LCMRL would in general be used during new method development, while the (simpler) MRL procedure would be used by individual laboratories to document compliance with a limit that had been identified by the LCMRL. The possibility of the MRL as a single lab quantitation limit procedure should be further considered, although it does not currently incorporate on-going verification.

IDE/IQE

The IDE and IQE procedures suffered from the large differences in performance from lab to lab. Unfortunately it appears likely that substantial differences between laboratory capability are to be expected, and this is likely to be even more apparent with very highly sensitive methods where the ultimate capability is controlled by the ability of the laboratory to avoid contamination and interferences rather than the sensitivity of the analytical instrumentation. Under well controlled conditions, the IDE/IQE can do a good job of modeling the precision of the analytical method across its range of applicability. Additionally, the IDE and IQE procedures failed to meet the study MQOs more frequently than other procedures because they estimate "lowest possible" quantitation limits based on MQO targets that are either more stringent (false positive error rate at L_C = 1%) RSD at the IQE20) or less stringent (false negative rate at the IDE) than those of the study. The IDE/IQE procedures should be used for method promulgation and in the evaluation of multi-laboratory performance because of its ability to assess MQO performance across the entire concentration range of the method, for multiple labs..

Consensus Group Procedure

The Consensus group procedure was not piloted. In most respects it can be considered as a more detailed and complex version of the ACIL procedure. Concepts from the Consensus group procedure should be used to inform modifications to the ACIL procedure in the light of the Pilot Study results.

LabQC Procedure

The LabQC procedure was not piloted. The basic concepts in the LabQC procedure are very similar to those in the ACIL procedure. As such the concepts from the LabQC procedure should be used to inform modifications to the ACIL procedure in the light of the Pilot Study results.

iv. Ruggedness Testing

If resources are available it is recommended that the modified ACIL procedure be piloted. It would also be beneficial to collect data for some of the more recent, highly sensitive technologies, for example methods 200.8 and 1631.

v.Lab Comments

ACIL Procedure

Most of the lab comments indicated that the ACIL procedure was straightforward to understand and apply. Several commenter's also indicated that they believed that the procedure was a considerable improvement over the current MDL procedure. Some commenter's noted that the determination of detection and quantitation limits was based on reagent water, and therefore might or might not apply to a real world matrix. Some suggestions fro improvement were made, these will be considered during the process of modification of the ACIL procedure based on Pilot Study results.

LCMRL

Most of the comments indicated that the LCMRL calculator was straightforward to apply. Some noted that the limits generated might not apply to real world matrices (similar to comments for the ACIL procedure). There were a few comments that the LCMRL procedure should mention the importance of non-consecutive analyses

III. Pilot Study Report Appendix: List of Tables

For Tables see Pilot Study Report Appendix.

Table Number Table Information

1. False Positive Rate Summary Stats, single-lab limits
2. False Positive Rate Summary Stats, interlab limits
3. Quantitation Limit MQO Summary Stats, single-lab limits, no outlier removal
4. Quantitation Limit MQO Summary Stats, single-lab limits, with outlier removal
5. Quantitation Limit MQO Summary Stats, interlab limits, no outlier removal
6. Quantitation Limit MQO Summary Stats, interlab limits, with outlier removal
7. False Positive Rates for Single-Lab Limits, no outlier removal
8. False Positive Rates for Single-Lab Limits, with outlier removal
9. False Positive Rates for Interlab Limits, no outlier removal
10. False Positive Rates for Interlab Limits, with outlier removal
11. MQO Characteristics for Method 300.0 Single-Lab Quantitation Limits, no outlier removal
12. MQO Characteristics for Method 335.4 Single-Lab Quantitation Limits, no outlier removal
13. MQO Characteristics for Method 200.7 Single-Lab Quantitation Limits, no outlier removal
14. MQO Characteristics for Method 608 Single-Lab Quantitation Limits, no outlier removal
15. MQO Characteristics for Method 625 Single-Lab Quantitation Limits, no outlier removal
16. MQO Characteristics for Method 300.0 Single-Lab Quantitation Limits, with outlier removal
17. MQO Characteristics for Method 335.4 Single-Lab Quantitation Limits, with outlier removal
18. MQO Characteristics for Method 200.7 Single-Lab Quantitation Limits, with outlier removal
19. MQO Characteristics for Method 608 Single-Lab Quantitation Limits, with outlier removal
20. MQO Characteristics for Method 625 Single-Lab Quantitation Limits, with outlier removal
21. MQO Characteristics for Method 300.0 Interlab Quantitation Limits, no outlier removal
22. MQO Characteristics for Method 335.4 Interlab Quantitation Limits, no outlier removal
23. MQO Characteristics for Method 200.7 Interlab Quantitation Limits, no outlier removal
24. MQO Characteristics for Method 608 Interlab Quantitation Limits, no outlier removal
25. MQO Characteristics for Method 625 Interlab Quantitation Limits, no outlier removal